(9) Final Report of 1 Jan 77- 3ø Dec 78

(6) Research on Voice Authentication.

(15) Contract N00014-77-C-0264 VW ARPA Order-3135

(10) John D. Markel
Principal Investigator

(12) 55

Speech Communications Research Laboratory, Inc.
800A Miramonte Drive
~~Santa Barbara~~, California 93109
Los Angeles

(11) Feb[redacted]1978

Accession For		
NTIS GRA&I		☒
DTIC TAB		□
Unannounced		□
Justification		
By Per Ltr. on file		
Distribution/		
Availability Codes		
Dist	Avail and/or Special	
A		

Subject: Final Report

Title: Research on Voice Authentication

Contract: N00014-77-C-0264

ARPA Order No: 3135

Contractor: Speech Communications Research Laboratory, Inc.
800A Miramonte Drive
Santa Barbara, CA 93109

Contract Dates: 1/1/77 - 12/30/78

Principal Investigator: J.D. Markel, Ph.D.

Summary of Research:

The purpose of this study was to develop voice authentication techniques which can be used over packet switched networks without the limitations of a single fixed reference phase. The results of this study are presented in the two manuscripts which have been or will be published (in February 1979) in the IEEE Acoustics, Speech and Signal Processing Journal.*

The results of this study were significant in several respects. First of all, the largest data base of controlled conditions but extemporaneous human speech in existance was developed for the project. Secondly, a real-time processing capability was developed for processing this data base of 40 hours of continuous speech. Finally, with the restriction of clean speech (not processed over a telephone or with noise), text-independent speaker recognition results nearly matching those for text-dependent studies were achieved by averaging over 30-40 second speech segments.

* 1) J.D. Markel, Beatrice T. Oshika, and A.H. Gray, Jr., Long-Term Feature Averaging for Speaker Recognition, IEEE Trans. Acoust., Speech, and Signal Proc., vol. ASSP-25, no. 4, August 1977.

2) J.D. Markel, and S.B. Davis, Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base, to be published in IEEE Trans. Acoust., Speech, and Signal Proc., February 1979.

Long-Term Feature Averaging for Speaker Recognition

JOHN D. MARKEL, MEMBER, IEEE, BEATRİCE T. OSHIKA, AND AUGUSTINE H. GRAY, JR., MEMBER, IEEE

Abstract—The potential benefits of long-term parameter averaging for speaker recognition were investigated. Parameters studied were pitch, gain, and reflection coefficients. Parameter variability was computed over various averaging lengths from one frame averaging (in effect, no averaging) to 1000 frame averaging (about 70 s of speech). It was demonstrated that the between-to-within speaker variance ratio, measured over several speakers, was significantly increased by performing long-term averaging of the parameter sets. The reflection coefficient averages for k_2 and k_6, respectively, were shown to produce the highest variance ratios.

I. Introduction

THERE have been several studies on the choice of acoustic features in speaker recognition tasks [14], [19], [22]. Average fundamental frequency has been found to be a useful discriminating feature [13], as have gain measurements [2], [10] and long-term speech spectra [4]-[6], [9]. Perceptual studies indicate that "there is at least weak evidence that a voice that is distinctive to listen to also has distinctive spectrographic patterns" [20], and that dimensions of "characteristic pitch" and "characteristic loudness" may be posited to differentiate among speakers [21]. These speaker characteristics can be distinguished from the acoustic cues which signal linguistic elements, e.g., phonemes or words. For example, the realization of the word "bit" by a female child is acoustically very different from the same word pronounced by an adult male, yet the words are generally understood to be equivalent while the speakers are clearly different. It appears, then, that listeners adapt to speakers' *voice* characteristics (as well as their *linguistic* characteristics).

All this suggests that there are long-term characteristics which can be used in text-independent speaker recognition tasks. Such characteristics include long-term averages related to fundamental frequency, gain, and spectral averages.

The motivation for long-term averaging in text-independent speaker recognition is based upon a result from statistical sampling theory.

We assume that $\{p(i)\}$ defines statistically independent, identically distributed samples of the parameter p with true mean μ_p and variance σ_p^2. (For example, $\{k_1(i)\}$ corresponds to the reflection coefficient k_1 samples for each analysis frame.) If $x = \langle p(i) \rangle$ defines a *feature* based upon long-term averaging of p, where

$$\langle p(i) \rangle = \frac{1}{L_v} \sum_{i=0}^{L_v-1} p(i), \qquad (1$$

and L_v is the number of voiced analysis frames used in the averaging, then the variance of x is given in terms of the origina

Manuscript received May 13, 1976; revised December 13, 1976 and March 15, 1977. This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Office of Naval Research under Contract N00014-73-C-0221.

J. D. Markel and B. T. Oshika are with the Speech Communications Research Laboratory, Inc., Santa Barbara, CA 93109.

A. H. Gray, Jr., is with the Speech Communications Research Laboratory, Inc., Santa Barbara, CA 93109 and the Department of Electrical Engineering and Computer Science, University of California, Santa Barbara, CA 93109.

Printed in U.S.A. Annals No. 708AS007

parameter variance σ_p^2 by

$$\sigma^2 [\langle p(i) \rangle] = \sigma_p^2 / L_v. \tag{2}$$

The sample variance as a function of L_v is an important figure of merit for a particular feature. For example, if the features are more tightly clustered together about the sample mean as L_v increases from $L_v = 1$ (no averaging), then the intraspeaker variability is decreased, and the parameters would be expected to result in higher performance in text-independent speaker recognition tasks. Although no "true mean or true variance" exists for real speech because of physiological variations in human speech, it is reasonable to assume that at least some convergence or clustering of parameters will occur with long-term averaging.

The purpose of this paper is to define several sets of potentially useful long-term features and then to investigate their statistical properties as a function of the averaging length L_v. In addition, discrimination tests are presented over a small homogeneous set of speakers to illustrate the potential benefits of long-term averaging for unconstrained text-independent speaker recognition.

II. Features

To discuss the applicability of long-term feature averaging in a quantitative manner, we have chosen three different feature sets as the basis for analysis. Some of these features reflect physiological characteristics more closely than others.

A. Fundamental Frequency Features

Due to physiological considerations such as the length and thickness of the vocal folds, and respiratory muscle patterns, the phonation of a particular vowel with "normal effort" may result in differing rates of vocal fold vibration (corresponding to the acoustical correlate of fundamental frequency) for different speakers. For example, a child will have a high fundamental frequency compared to an adult because of the child's smaller vocal folds.

Although fundamental frequency, along with intensity and duration, is a controllable attribute of stress and intonation which may vary widely, each person appears to have a mean fundamental frequency value which, if averaged over a sufficiently long period of time, is relatively constant over a reasonable time span and is independent of linguistic content [8].

In addition, the standard deviation of the fundamental frequency over a long interval of time may carry important speaker-dependent information. For example, if the speaker is judged to be a monotone speaker, then the standard deviation would be expected to be relatively small. However, if the speaker is thought to be an "expressive" or "forceful" speaker, it would be expected to be relatively large.

B. A Gain Feature

It seems reasonable to assume that one of the characteristics that contributes to a speaker's identity is the amount of intensity or gain variation in his speech over time. Subjectively, the amount of gain variation is possibly correlated with the perception of "dynamic" versus "flat" voices. The actual gain variation is also a function of phonetic content, word and phrase stress, and discourse context. For example, for a constant subglottal pressure, the acoustical output energy for an /a/ is about 5 dB greater than for a /u/. Also, a larger gain variation would be expected with an exclamatory as opposed to a normal declarative sentence. Our assumption is that, over a sufficiently long interval of speech, gain variation can be considered part of the individual speaker's characteristics. That is, a speaker who is judged overall to be an "emphatic" speaker will have larger gain variation than one who is judged to have a usually monotonous voice.

In the measurement of gain variation, it is very important that results be only a function of speaker characteristics and not absolute system gain. Furthermore, because of the distinctly different production mechanism between voiced and unvoiced speech, it is desirable to measure the gain variations only during voiced speech. A normalized gain variation which satisfies desired physical properties is now defined. If $R(n)$ defines the energy of N speech samples $\{s(l)\}$ in frame n, then

$$R(n) = \sum_{l=0}^{N-1} s^2(l). \tag{3}$$

The sample mean and sample variance of $R(n)$ over L_v voiced frames is then defined by

$$\overline{R} = \langle R(n) \rangle \tag{4}$$

and

$$\sigma_R^2 = \langle (R(n) - \overline{R})^2 \rangle \tag{5}$$

where $\langle \cdot \rangle$ will be used throughout to denote averaging over L_v voiced frames. The normalized gain variation δ is then defined by

$$\delta = \sigma_R / \overline{R}. \tag{6}$$

If the overall system gain is changed by a constant value, δ is unaffected. Furthermore, δ is nonnegative with $\delta = 0$ only when $\sigma_R = 0$. Physically, $\delta = 0$ means that the speech envelope (more precisely, the frame energy) is unchanged over the complete range of voiced speech analyzed.

C. Spectral Features

It is well established in the literature that one of the acoustical features that tends to differentiate one particular speaker from another during voiced speech production is the glottal sound source shape [15].

Although the spectral slope of a single glottal pulse can vary over a wide range from nearly whispered speech to very intense vocal effort, for normal conversational speech it is expected that an average glottal source spectrum could be obtained over a relatively long interval of speech that would have relatively small intraspeaker variability.

Unfortunately, glottal volume velocity waveform estimation from speech is a nontrivial task [7], [12], [16]. A more direct method for automatic real-time analysis is to use a parameter set that is related to the smooth characteristics of the spectrum, which is independent of fundamental frequency or gain. With linear prediction analysis, obvious possibilities are filter coefficients, reflection coefficients, or log area functions. Sambur [17] compared these coefficients in a speech recognition experiment and decided to make use of the reflection coefficients. Although reflection coefficients are nonlinearly related to the more physically meaningful smooth-spectral and log-spectral model from linear prediction analysis, there is ample evidence that they do contain important speaker-dependent information that is not contained in fundamental frequency- or gain-related parameters. For example, in the case of a first-order filter, $M = 1$, a smooth spectral model can be physically and mathematically related to the first reflection coefficient. This model [11, p. 139] has a spectral flatness given by

$$\Xi(1/A) = 1 - k_1^2. \tag{7}$$

If the speech sample being analyzed has a nearly flat spectral trend, k_1 approaches zero and the spectral flatness approaches unity. As the spectral slope increases negatively, k_1 approaches -1 and the spectral flatness approaches zero.

Based upon the spectral matching properties of linear prediction [11, p. 134], we would assume that preemphasis of the data would be beneficial since the reflection coefficients would then carry more information about the spectral structure at higher frequencies.

It would also seem reasonable that if long-term, spectrally related features are desired which minimize intravariability, only voiced speech should be analyzed. Substantial differences exist in the physiological mechanisms which produce voiced and unvoiced sounds. Since the excitation for unvoiced speech is generally assumed to have a flat spectrum, the difference in spectral slope between voiced and unvoiced sounds may be on the order of 8-16 dB. With only voiced sounds, some variation will still occur since different articulator positions will cause variations on the acoustic loading at the glottis, affecting the glottal source shape. This variation, however, is expected to be substantially less than that due to glottal source variations in voiced-unvoiced speech production.

D. *Summary of Feature Definitions*

As *features* we study the following.

1) F_0 average

$$x_1 = \overline{F}_0 = \langle F_0(n) \rangle. \tag{8}$$

2) Standard deviation of F_0

$$x_2 = \sigma_{F_0} = \langle [F_0(n) - \overline{F}_0]^2 \rangle^{1/2}. \tag{9}$$

3) Sample gain variation

$$x_3 = \sigma_R / \overline{R} \tag{10}$$

where

$$\overline{R} = \langle R(n) \rangle \tag{11}$$

and

$$\sigma_R = \langle [R(n) - \overline{R}]^2 \rangle^{1/2}. \tag{12}$$

4) Spectrally related features (reflection coefficient averages)

$$x_{i+3} = \langle k_i(n) \rangle \quad \text{for} \quad i = 1, 2, \cdots, M. \tag{13}$$

The *feature vector* x is defined by

$$x^T = [x_1 x_2 \cdots \ x_{3+M}] \tag{14}$$

where T denotes transpose.

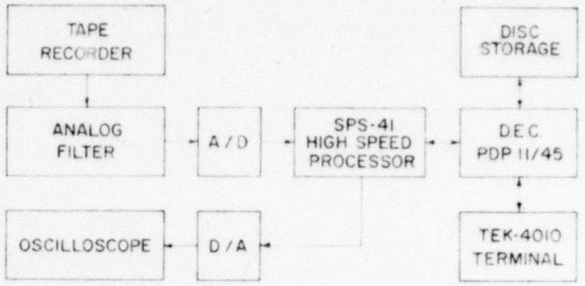


Fig. 1. Block diagram of system for processing speech.

III. Procedures

A. *Data*

The data used for the analysis were obtained during interviews of four speakers. Each interview was then edited so that only the interviewees' voices remained. The total duration of each edited interview (including pauses) was typically 15-18 min. The total data base used for this study was approximately one hour in duration. No special precautions or recording conditions were imposed on the experiment. Interviews were conducted in normal room environments with a dynamic microphone and an audio tape recorder. So that a small number of speakers could be used with some generality in extrapolating results, a homogeneous population of four male speakers was chosen, each having somewhat similar speech characteristics and relatively narrow fundamental frequency ranges. Histograms of the raw nonaveraged fundamental frequency values showed substantial overlap among the four speakers.

B. *Digital Processing of Data*

The audio tape was digitally processed using the system shown in Fig. 1. Each test segment was recorded onto a disk using conventional procedures. A novel part of the procedure is based upon the use of a high-speed signal processor and oscilloscope (for visual feedback during processing). Using an array-processing software system, it is possible to process the data in real time at a 50 Hz analysis frame rate from a Fortran environment. Processing includes modified cepstral pitch period and voicing detection, gain calculation, linear prediction analysis for reflection coefficients, and a running mean and mean-square computation of these parameters.

The procedure for generating output feature vectors to be used in the statistical analysis is shown in Fig. 2. A counter for frame n is incremented and one frame of speech is analyzed. The parameters used are: sampling frequency f_s = 6.5 kHz, number of analysis coefficients $M = 10$, number of samples for reflection coefficient computation = 128, and the number of samples for F_0, and gain parameter analysis = 256 (40 ms). The analysis frame rate is 50 Hz. Preemphasis of the speech data is applied using a differencer, $1 - z^{-1}$.

Fundamental frequency estimation is performed with a modified cepstral technique. After the spectrum has been computed, a symmetrical window function is applied that smoothly tapers from unity at 1000 Hz to zero from 1500 Hz to $f_s/2$. This simple modification resolves most of the voicing problems one obtains with the usual cepstral analysis method since only

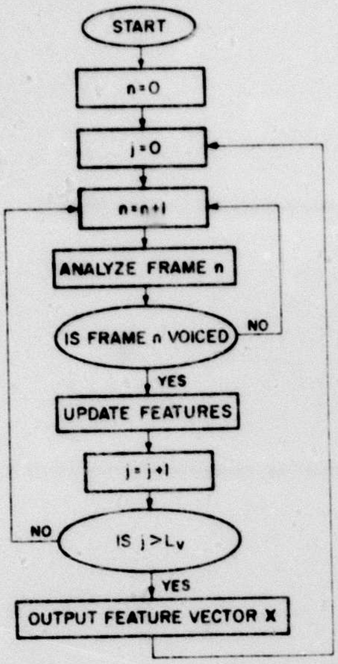


Fig. 2. Procedure for generating output feature vectors.

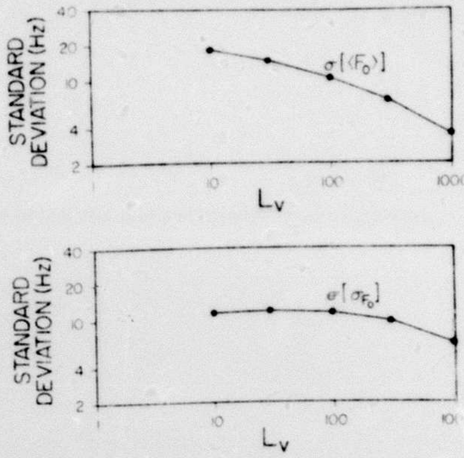


Fig. 3. Standard deviation of F_0 related features as a function of the number of voiced frames L_v.

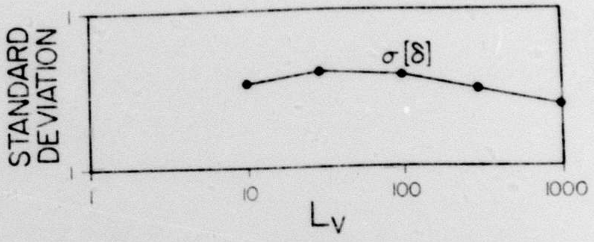


Fig. 4. Standard deviation of gain related features as a function of the number of voiced frames L_v.

the most consistent region of harmonic structure is used [3]. Two frames of delay are included in the system so that some amount of error detection and correction can be applied in the pitch period estimation. One additional test has been found necessary for obtaining meaningful feature vectors. A max (F_0) and min (F_0) value are chosen for the speaker being analyzed to ensure against gross errors causing the fundamental frequency features from being dramatically affected. If min $(F_0) < F_0 <$ max (F_0), the frame is judged to be voiced and the long-term averages are updated. The frame counter is incremented and if $l > L_v$, the resultant features vector x is output to disk, l is reset to zero, and analysis then continues.

IV. Experiments

A. Experiment 1–Statistical Variation as a Function of L_v

The complete edited audio tape for speaker D (approximately 18 min in duration) was analyzed to extract long-term averaged feature vectors for several L_v conditions. As a time reference, $L_v = 1000$ corresponds to approximately 70 s. The total number of vector samples obtained is approximately inversely proportional to L_v.

The unbiased variance estimate of the feature $x = \langle p(i) \rangle$ based upon the speech parameter p is

$$\sigma^2(x) = \frac{1}{L_f - 1} \sum_{i=0}^{L_f - 1} [\langle p(i) \rangle - \bar{x}]^2 \tag{15}$$

where

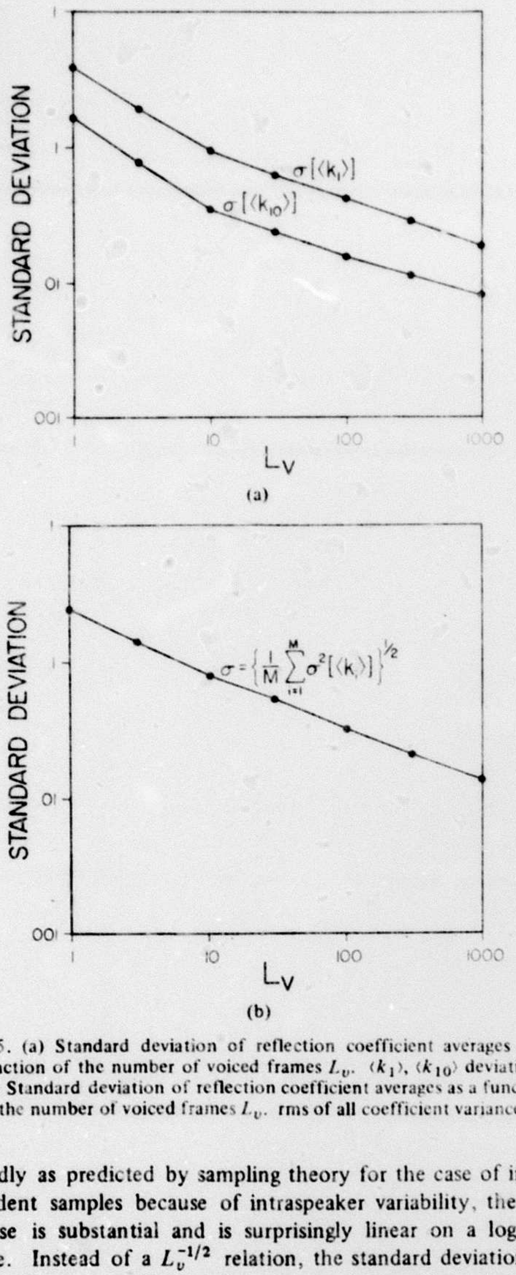
$$\bar{x} = \frac{1}{L_f} \sum_{i=0}^{L_f - 1} \langle p(i) \rangle. \tag{16}$$

Each $p(i)$ explicitly denotes an individual feature, and L_f is the number of feature vectors obtained over the total speech duration. Note that L_f is actually a function of L_v since the total duration is fixed. The sample mean x is thus independent of L_v except for sampling variation in the real-time analysis because it is not possible to start analysis at precisely the same location on the audio tape when L_v is changed. The true variance σ_p^2 is estimated from $\sigma_p^2 = \sigma^2(x)$ with $L_v = 1$. Features which themselves are based on variances (such as $x_2 = \sigma_{F_0}$ and $x_3 = \sigma_R / \bar{R}$) do not allow for a true variance estimate. The sample standard deviations of the fundamental frequency-related features are shown in Fig. 3 as a function of L_v. The estimated standard deviation about the long-term fundamental frequency averages is reduced from about 18 Hz for $L_v = 10$ to about 6 Hz for $L_v = 1000$. These values are somewhat higher than the long-term F_0 averages reported by Horii [8]. However, this experiment is based upon unconstrained conversational speech, whereas Horii's experiment was based upon a reading of the "Rainbow Passage."

The estimated standard deviation of the $x_2 = \sigma_{F_0}$ feature is surprisingly constant, until at least on the order of 7-10 s of speech ($L_v > 100$) have been analyzed. Increasing L_v from 100 to 1000 decreases $\sigma(x_2)$ from 12-6 Hz.

The variability of the gain variation feature, $\sigma(x_3)$, as shown in Fig. 4, follows a similar pattern. This particular feature appears to be a very weak function of L_v.

In Fig. 5(a), the estimated standard deviation of the $\langle k_1 \rangle$ and $\langle k_{10} \rangle$ features for speaker D are shown as being representative of the reflection coefficient feature set characteristics. Although the estimated standard deviation does not decrease as



(a)

(b)

Fig. 5. (a) Standard deviation of reflection coefficient averages as a function of the number of voiced frames L_v. $\langle k_1\rangle$, $\langle k_{10}\rangle$ deviations. (b) Standard deviation of reflection coefficient averages as a function of the number of voiced frames L_v. rms of all coefficient variances.

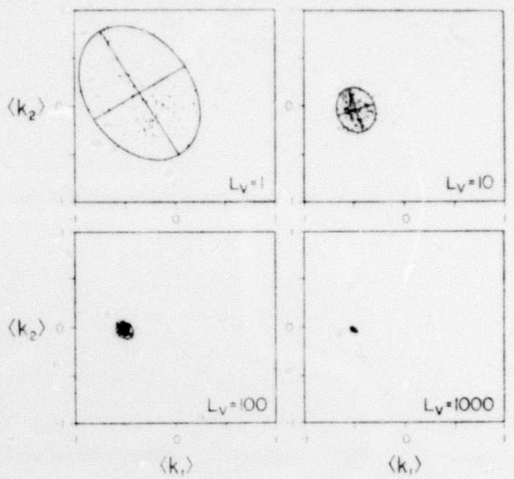


Fig. 6. Scatter plots of $\langle k_1\rangle$, $\langle k_2\rangle$ features for different L_v with two-sigma ellipses and principal axes.

rapidly as predicted by sampling theory for the case of independent samples because of intraspeaker variability, the decrease is substantial and is surprisingly linear on a log-log scale. Instead of a $L_v^{-1/2}$ relation, the standard deviation of the reflection coefficient features appears to approximately decrease proportionally to a $L_v^{-1/3}$ model beyond $L_v = 10$.

The rms deviation over all $\langle k_i\rangle$ averages is shown in Fig. 5(b). Over a range of L_v from 10 to 1000, the $L_v^{-1/3}$ model is still seen to be very accurate for predicting the decrease in reflection coefficient feature parameter variation as L_v is increased. The measured exponent value is certainly dependent upon the particular speaker. However, it appears to vary only slightly from the model discussed for the several other speaker measurements.

The estimate of the true variance for the k_1, k_{10}, and overall parameter variance is also shown in Fig. 5(a) and (b) at $L_v = 1$.

A second way of qualitatively showing the effect of long-term averaging is to show two-dimensional scatter diagrams for various values of L_v. Fig. 6 shows a scatter plot of $\langle k_2\rangle$ versus $\langle k_1\rangle$ samples. Each point is based upon L_v samples from the edited audio tape for speaker D. Shown with the data are two-sigma ellipses with the principal axes. A dramatic decrease in the dispersion of the data is seen as L_v increases.

B. Experiment 2–Discrimination as a Function of L_v

The approach taken here is to investigate the effectiveness of long-term averaging for speaker recognition using the ratio of the between-speaker variance and the within-speaker variance, without specifying particular speaker recognition experiments. Since the mathematics of this procedure (Fisher discriminant method) is discussed elsewhere [1], only the necessary details will be summarized below.

A within-speaker covariance matrix $\boldsymbol{W}$ is computed, and then a normalized between-speaker covariance matrix $\boldsymbol{B}'$ is found in terms of the matrix $\boldsymbol{B}$ of Bricker *et al.* [1] from

$$\boldsymbol{B}' = \boldsymbol{B}/L_f \tag{17}$$

where L_f is the number of feature vectors. The normalization is included so that B' will depend only upon the sample means, not upon the number of feature vectors. Eigenvalues and eigenvectors of the equation

$$\boldsymbol{B}'\boldsymbol{\phi}_k = \lambda_k \boldsymbol{W}\boldsymbol{\phi}_k \tag{18}$$

are then obtained. The eigenvalues are ordered from highest to lowest, and as the number of speakers, four, is less than the number of features, thirteen, all but the first three eigenvalues are zero [1]:

$$\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \geqslant \lambda_4 = \lambda_5 = \cdots = \lambda_{13} = 0. \tag{19}$$

A new coordinate system is defined using the eigenvectors of (18) as base vectors, so that the new coordinate vector $\boldsymbol{y}$ is related to $\boldsymbol{x}$ through the linear transformation

$$\boldsymbol{y} = \boldsymbol{\phi}^T \boldsymbol{x} \tag{20}$$

where $\boldsymbol{\phi}$ is the matrix whose columns are the eigenvectors of (18). The eigenvalues of (18) represent the variance ratios in the directions of the eigenvectors with λ_1 being the maximum

TABLE I
VARIANCE RATIOS OF LONG-TERM AVERAGE FEATURE SET FOR L_v = 100 AND L_v = 1000

FEATURES	VARIANCE RATIO	
	$L_V = 100$	$L_V = 1000$
$x_1 = \langle \bar{F}_0 \rangle$	0.332	2.321
$x_2 = \langle \sigma_{F_0} \rangle$	0.004	0.043
$x_3 = \langle \delta \rangle$	0.119	0.329
$x_4 = \langle k_1 \rangle$	0.081	0.305
$x_5 = \langle k_2 \rangle$	2.721	16.118
$x_6 = \langle k_3 \rangle$	0.221	1.216
$x_7 = \langle k_4 \rangle$	0.367	2.023
$x_8 = \langle k_5 \rangle$	0.307	2.002
$x_9 = \langle k_6 \rangle$	2.315	11.452
$x_{10} = \langle k_7 \rangle$	0.155	0.650
$x_{11} = \langle k_8 \rangle$	0.511	2.591
$x_{12} = \langle k_9 \rangle$	0.185	0.977
$x_{13} = \langle k_{10} \rangle$	0.403	0.978

TABLE II
VARIANCE RATIOS OF TRANSFORMED LONG-TERM AVERAGE FEATURE SET FOR L_v = 100 AND L_v = 1000

TRANSFORMED FEATURES	VARIANCE RATIO	
	$L_V = 100$	$L_V = 1000$
y_1	10.959	115.368
y_2	0.972	7.956
y_3	0.393	1.730
y_4	0	0
⋮	⋮	⋮
y_{13}	0	0

variance ratio, λ_2 being the next largest (in a direction orthogonal to Φ_1), etc. Variance ratios can also be computed in the original coordinate system as a method for measuring relative effectiveness of features.

Tables I and II show the variance ratios in the original and transformed coordinate systems, respectively, for $L_v = 100$ and $L_v = 1000$. Except for the fact that parameter correlation is not taken into account, the variance ratio values can be taken as quantitative measures of the original parameter's effectiveness in speech recognition. For example, we see that σ_{F_0} provides very little discrimination among speakers, whereas $\langle k_2 \rangle$ appears to provide the maximum discrimination among speakers over all parameters. It is seen that the first dimension in the new coordinate system results in a substantially increased variance ratio.

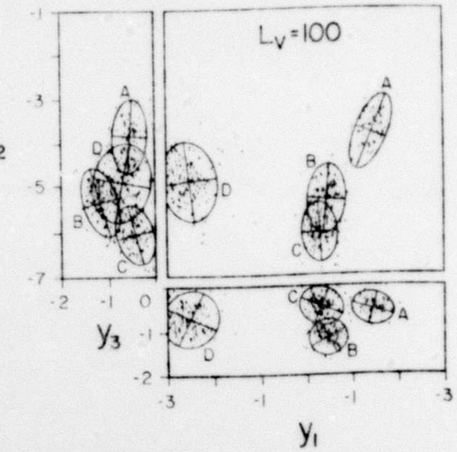


Fig. 7. Scatter plots for speakers *A-D* along first three Fisher discriminant dimensions (L_v = 100).

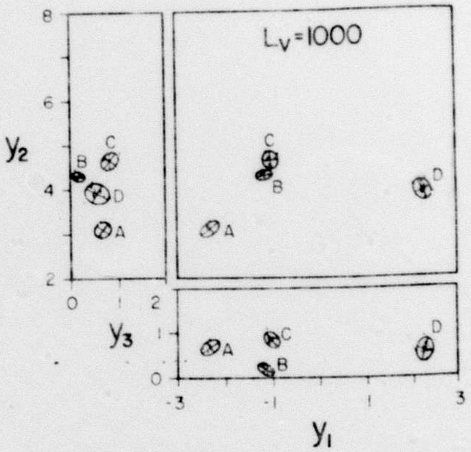


Fig. 8. Scatter plots for speakers *A-D* along first three Fisher discriminant dimensions (L_v = 1000).

Two-dimensional scatter plots of the first three transformed dimensions are shown in Fig. 7 for L_v = 100 and in Fig. 8 for L_v = 1000. The results are based upon the four speakers *A-D*. Also shown are two-sigma ellipses and the principal axes for each speaker distribution. In Fig. 7 it is seen that *A* and *D* are essentially uniquely separated from *B* and *C* in at least one plane ($y_1 - y_2$). A relatively large overlap does occur, however, for *B* and *C* in all planes. A cursory comparison of Fig. 7 and the relative sizes of clusters in Fig. 6 will illustrate that substantial benefits in discriminating against different speakers have been obtained over using no averaging (L_v = 1) or very limited amounts of averaging (L_v = 10).

In Fig. 8, it is seen that by performing long-term averaging with L_v = 1000, perfect discrimination is obtained, in this instance based upon only a two-dimensional transformed feature representation.

The variance ratios for the input feature variables are shown in Table I for L_v = 100 and L_v = 1000. If the variables were statistically independent, these ratios would differ by a multiplicative factor of 10 rather than the smaller factors indicated in the figure. The ordering of the features in terms of variance ratios is of some interest. The fifth feature, k_2, clearly shows the largest variance ratio, with the ninth feature, k_6, the next largest. These coefficients correspond to the coefficients for the highest power of z^{-1} in the models of order 2 and 6 found from linear prediction analysis. The two-pole model has been used in earlier recognition tasks [18].

The variance ratios for the first three features, fundamental frequency, its standard deviation, and sample gain variation, are smaller than what one might expect from intuition. Part of the reason may lie in the fact that the speakers were chosen to have similar fundamental frequency ranges.

The variance ratios for the new coordinate system, the eigenvalues of (18), are shown in Table II. From these ratios and the scatter diagrams of Fig. 8, it can be seen that very clear separation of the speakers is indicated for the long-term average case of $L_v = 1000$ by using only the first two coordinates, y_1 and y_2, in the direction of the eigenvectors $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$.

V. Discussion

A. Parameter Variability Over Days, Weeks, Etc.

This initial study has been restricted to the study of long-term averages taken from one session. This is probably the reason why the standard deviation of the long-term averages tends to have a monotonically decreasing behavior. Although some amount of intraspeaker variability is reflected in the data, additional variability will occur when results are obtained from sessions separated by days, weeks, or months later. In several studies over linguistically constrained units, this effect has been shown to be severe beyond several months for short text-dependent segments [4]. A large data base extending over several months is now being generated for studying these effects in conversational speech.

B. Accuracy of Voicing Decisions

Since all long-term statistics are made only during voicing, it is very important to know that realistic voicing decisions are made. Spectral slope and normalized gain variation are direct computations requiring no decisions (except for voicing) and are, therefore, very robust.

If the threshold setting for voicing and pitch period detection is set too high or too low, the effect can be catastrophic. At one extreme, if the voicing threshold is too high, very few frames will be included in the statistics as being voiced (although they will be very reliable estimates) and, furthermore, transitions in which considerable fundamental frequency variations may occur are likely to be missed, causing the measured fundamental frequency standard deviation to be unrealistically small.

At the other extreme, if the threshold is too low, there will be a tendency to define fundamental frequencies near the maximum allowable frequency (minimum pitch period) (near 400 Hz) during actual voiced speech and at random values throughout the rest of the allowable range during unvoiced speech. Although a pitch period and voicing decision program with several frames of delay is used for error detection and correction, it is essentially impossible to separate accurate estimates from gross errors beyond some reasonable threshold.

C. Assumptions Versus Experimental Results

It was assumed that $\langle F_0 \rangle$ carries important speaker information. The $\sigma[\langle F_0 \rangle]$ versus L_v graph in Fig. 3 showed a significant monotonic decrease as L_v was increased. In addition, the variance ratio was relatively high (even though speakers were purposely chosen with similar fundamental frequency ranges). Therefore, this assumption appears valid. The assumption that $\sigma[\sigma_{F_0}]$ is meaningful does not appear to be true for conversational speech. The variance ratio for this feature is extremely small. This result contradicts that shown by Mead [13], where the use of the first through the fourth moments of F_0 and of the first four differences of F_0 (resulting in 20 features) was suggested. Our experience indicates that unless hand-marked or hand-corrected F_0 contours are used, very significant biases in results can occur because of very few gross errors in F_0 estimation. Higher order differences and moments only magnify these biases.

The standard deviation of the gain deviation feature as a function of L_v shows a weak relationship to expectations from statistical sampling theory. In addition, the variance ratios for the gain deviation feature are relatively small. Although some discrimination is obtained, what we have seen is that not only is there substantial intraspeaker variability for this parameter, but that, in addition, considerable overlap in the gain feature values occurs between speakers. Other measures of fundamental frequency and gain variations may prove to be more useful than the ones used here, which are essentially based upon root mean squares taken about the averages. One possibility is the use of the ratio of geometric and arithmetic means as used in evaluating spectral flatness [11].

The long-term averages of the reflection coefficients as a set appear to be the most significant features for speaker recognition. Not only does the standard deviation of the long-term averages show a substantial decrease as a function of L_v, but in addition, the variance ratios are seen to be relatively large for most of the parameters.

D. Observations on Reflection Coefficient Averaging

Although $\sigma(\langle k_{10} \rangle) < \sigma(\langle k_1 \rangle)$ for all L_v, in Fig. 5(a), one should not be misled into thinking that $\langle k_{10} \rangle$ is a better feature for speaker recognition. This result occurs because k_1 inherently has a larger standard deviation than k_{10} ($\langle k_1 \rangle = k_1$ for $L_v = 1$). The important fact to note is that whatever the parameter deviation is without averaging, due to either linguistic content or intraspeaker variability, it decreases as $L_v^{-\alpha}$ where $\frac{1}{3} \leqslant \alpha \leqslant \frac{1}{2}$ when long-term averaging is applied.

In a recent paper [17], the use of orthogonal linear prediction parameters for use in text-independent speaker recognition studies was suggested. Although very high recognition scores were shown using the orthogonal linear prediction parameters, we would suggest that substantial reduction in scores would occur if unconstrained data bases as described here were used. Whatever scores are obtained using, in effect, $L_v = 1$, our results qualitatively indicate that substantial improvements could occur by incorporating long-term averaging.

Each orthogonal parameter was obtained from a linear combination of all reflection coefficients as

$$\Phi_i = \sum_{j=1}^{M} c_{ij} k_j \tag{21}$$

where the c_{ij} terms were obtained from a principal component analysis. The averaged parameters would then be

$$\langle \Phi_i \rangle = \sum_{j=1}^{M} c_{ij} \langle k_j \rangle. \tag{22}$$

Although Fig. 6 shows only the dispersion characteristics for $\langle k_1 \rangle$ and $\langle k_2 \rangle$, similar characteristics are obtained for all the coefficients. The amount of data dispersion will be primarily due to the value of L_v, not the fact that a linear combination of the k_i terms (or the $\langle k_i \rangle$ terms) has been obtained.

E. Computational Considerations

Studies of this type place a premium on the available processing speed of the computer system. It became clear early in the study that small- or medium-scale computer capability was insufficient. For example, the analysis method described runs in approximately 100 times real time if all operations are implemented only on the PDP-11 system. The relatively small data base of speakers for this study would have required over 100 hours of processing time.

Except for the nontrivial costs in software development, we have found that attaching a high-speed processor to the main computer system provides a very economical solution to the requirements for real-time processing.

VI. Summary

The properties of long-term feature averaging for three sets of fundamental frequency related, gain related, and spectrally related parameters have been investigated. Based upon the Fisher discriminant method, the rank ordering of the parameter sets in importance was shown to be spectral, fundamental frequency, and then gain. It was also shown that over a long duration from $L_v = 10$ to $L_v = 1000$, the standard deviation of the sample means of the reflection coefficient vectors decreased proportionally to $L_v^{-1/3}$.

A small number of speakers with relatively homogeneous characteristics was used to illustrate the effects of long-term averaging. The data base was of nontrivial duration, somewhat greater than one hour in length. Furthermore, the text was unconstrained conversational speech, recorded under normal room noise conditions. Analysis was performed in real time with a high-speed signal processor.

Presently, other spectral representation methods are being investigated and a data base is being developed for performing text-independent speaker recognition tests without any linguistic or structural constraints.

References

[1] P. D. Bricker, R. Gnanadesikan, M. V. Mathews, S. Pruzansky, P. A. Tukey, K. W. Wachter, and J. L. Warner, "Statistical techniques for talker identification," *Bell Syst. Tech. J.*, vol. 50, pp. 1427–1454, 1971.

[2] G. R. Doddington, "A method of speaker verification," Ph.D. dissertation, Univ. of Wisconsin, Madison, 1970.

[3] O. Fujimura, "An approximation to voice aperiodicity," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 68–72, 1968.

[4] S. Furui, "An analysis of long-term variation of feature parameters of speech and its application to talker recognition," *Electron. Commun. Jap.*, vol. 57-A, pp. 34–42, 1974.

[5] S. Furui and F. Itakura, "Talker recognition by statistical features of speech sounds," *Electron. Commun. Jap.*, vol. 56-A, pp. 62–71, 1973.

[6] S. Furui, F. Itakura, and S. Saito, "Talker recognition by long-time averaged speech spectrum," *Electron. Commun. Jap.*, vol. 55-A, pp. 54–61, 1972.

[7] J. N. Holmes, "Low-frequency phase distortion of speech recording," *J. Acoust. Soc. Amer.*, vol. 58, pp. 747–749, 1975.

[8] Y. Horii, "Some statistical characteristics of voice fundamental frequency," *J. Speech Hearing Res.*, vol. 18, pp. 192–201, 1975.

[9] U. Kosiel, "Statistical analysis of speaker-dependent differences in the long-term average spectrum of Polish speech," in *Speech Analysis and Perception*, vol. 3, W. Jassem, Ed., Polish Academy of Sciences. Warsaw, Poland: PWN-Polish Scientific Publishers, 1973, pp. 117–120.

[10] R. C. Lummis, "Speaker verification by computer using speech intensity for temporal registration," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 80–89, 1973.

[11] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. Berlin, Heidelberg, New York: Springer-Verlag, 1976.

[12] J. D. Markel and D. Wong, "Considerations in the estimation of the glottal volume velocity waveforms," submitted to *IEEE Trans. Acoust., Speech, Signal Processing*.

[13] K. O. Mead, "Identification of speakers from fundamental frequency contours in conversational speech," Joint Speech Res. Unit, Rep. 1002, 1974.

[14] W. S. Mohns, "Statistical feature evaluation in speaker identification," Ph.D. dissertation, North Carolina State Univ., Raleigh, 1969.

[15] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Amer.*, vol. 49, pp. 583–590, 1970.

[16] M. Rothenberg, "A new inverse-filter technique for deriving the glottal air flow waveform during voicing," *J. Acoust. Soc. Amer.*, vol. 53, pp. 1632–1645, 1973.

[17] M. R. Sambur, "Speaker recognition using orthogonal linear prediction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 283–289, Aug. 1976.

[18] M. R. Sambur and L. R. Rabiner, "A speaker independent digit recognition system," *Bell Syst. Tech. J.*, vol. 54, pp. 81–102, 1975.

[19] K. N. Stevens, "Sources of inter- and intra-speaker variability in acoustic properties of speech sounds," in *Proc. 7th Int. Congr. of Phonetic Sciences*, A. Rigault and R. Charbonneau, Ed. The Hague: Mouton, 1972.

[20] K. N. Stevens, C. E. Williams, J. R. Carbonell, and B. Woods, "Speaker authentication and identification: A comparison of spectrographic and auditory presentation of speech material," *J. Acoust. Soc. Amer.*, vol. 44, pp. 1596–1607, 1968.

[21] W. D. Voiers, "Perceptual bases of speaker identity," *J. Acoust. Soc. Amer.*, vol. 36, pp. 1965–1973, 1964.

[22] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Amer.*, vol. 51, pp. 2044–2056, 1972.

Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base

John D. Markel and Steven B. Davis

May 1978

Signal Technology, Inc.
15 W. De La Guerra Street
Santa Barbara, CA 93101

Part of this work was performed while the authors were at the Speech Communications Research Laboratory, Inc., 800A Miramonte Drive, Santa Barbara, CA 93109.

This work was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Office of Naval Research under Contract N00014-77-C-0264

Abstract

Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base

John D. Markel and Steven B. Davis

A very large data base consisting of over thirty-six hours of unconstrained extemporaneous speech, from seventeen speakers, recorded over a period of more than three months, has been analyzed to determine the effectiveness of long-term average features for speaker recognition. Results are shown to be strongly dependent on the voiced speech averaging interval L_v. Monotonic increases in the probability of correct identification and monotonic decreases in the equal error probability for speaker verification were obtained as L_v increased, even with substantial time periods between successive sessions. For L_v corresponding to approximately thirty-nine seconds of speech, text-independent results (no linguistic constraints embedded into the data base) of 98.05% for speaker identification and 4.25% for equal error speaker verification were obtained.

I. Introduction

In recent years, there has been an increasing interest in computer-based techniques for text-independent speaker recognition (1-6). Recognition is used here to encompass both speaker identification and verification (7). The term "text-independent" has been used in several different contexts. For example, Atal (1) has used the term in the sense of choosing independent randomized test frames from a single sentence to use against the remaining frames as a reference set. Sambur (4) has used the term in an experiment where the sentences in the test set were different from those in the reference set, even though each speaker read precisely the same list of sentences.

Although useful insight has been gained by these approaches, they were linguistically constrained. In many practical situations, where text-independent speaker recognition is desired, there typically will be no control over the speech being tested. As Beek, Neuberg and Hodge (8) have pointed out, text-independent speaker identification can overcome problems which may arise if the speaker is uncooperative, and there is a great interest for speaker identification over communications channels, which have no linguistic constraints. Furthermore, there may be days to weeks of separation between reference and test sessions.

Several other studies (2,3,5,6) have analyzed data with varying amounts of linguistic constraints. Li and Walker (2) used thirty seconds of speech read from the rainbow passage (9) recorded once by twenty-two male speakers and twice by an additional eight male speakers. They did not specify the number of days separating the recordings. They demonstrated that distances among spectral correlation matrices could be used to compare inter-speaker and intra-speaker differences. However, the same text was used for all tests, which could be interpreted as a linguistic constraint.

Hunt, Yates and Bridle (6) used approximately six two- to three-minute long FM radio weather forecasts from each of eleven male and two female speakers. Each forecast was divided into twenty- or thirty-second intervals and long-term fundamental frequency and cepstral coefficient features were computed for twenty-millisecond sequential frames in each interval. They did not specify the number of days between successive forecasts by the same speaker. Using Fisher discriminant analysis (10), they achieved 89% correct speaker identification with independent test and reference sets. However, the speakers read text with some effort at uniformity between sessions, which could also be interpreted as a linguistic constraint.

In a preliminary study, Markel, Oshika and Gray (5) used one fifteen- to eighteen-minute interview from each of four male speakers with somewhat similar speech characteristics. The interviews were recorded with an audio tape recorder in a normal room environment. Long-term fundamental frequency, gain and reflection coefficient features were computed for every 1000 sequential voiced frames (twenty-millisecond windows per frame, fifty frames per second) in each interview. Using the same Fisher discriminant analysis (10) as Hunt et al. to transform the data, they achieved perfect discrimination among the four speakers. These recorded interviews were considered to be free of linguistic constraints. However, the data were insufficient to obtain statistically significant results, and with only one session per speaker, there was no analysis of speaker characteristics over time.

The purpose of this paper is to present results from experiments in speaker recognition where there were no linguistic constraints on the speech content (other than the ones implied when the speaker is cooperative, and English is used). In comparison with the previous study (5), results are presented for a larger number of speakers, for multiple sessions from each speaker, and for a greater number of features. Furthermore, the effects of time between recording sessions are studied. For practical implementation, only parameters obtained from the analysis portion of a linear prediction vocoder (fundamental frequency, gain and reflection coefficients) were used. (Beek et al. (8) have stated that the reflection coefficients are currently favored for all-digital narrowband communications systems.) This study shows that if these parameters are averaged over sufficiently long intervals of time, such as thirty seconds or more, the features obtained are essentially free of linguistic constraint, and speaker recognition performance is comparable with some text-dependent speaker recognition experiments. The linguistic results agree with Li and Walker (2), who used a smaller data base; long-term speech features are relatively stable after thirty seconds. Furthermore, this study shows that if the averaging interval is too short, speaker recognition performance is unacceptable with linguistically unconstrained extemporaneous speech. In addition, the importance of having a time-spaced reference set of sufficient size is demonstrated.

II. Data Base and Processing Methodology

A data base was collected by recording 170 fifteen-minute interviews from eleven male and six female speakers. There were ten sessions per speaker, with each session separated by a minimum of one week. Generally, the successive sessions were obtained within two to three weeks. One exceptional separation between successive sessions was fourteen weeks.

All sessions were recorded on a Tandberg 9000X two-track tape recorder at a recording speed of 7.5 ips. One track was used to record the interviewer and the other track was used to record the speaker. The speaker was recorded with a B and K half-inch condenser microphone and amplifier system in an IAC sound room equipped with a window. The interviewer was recorded with a conventional dynamic microphone outside of the sound room. Two-way communication was established using headphones.

Each session began with the speaker reciting his/her name, a password, a word list and the first paragraph of the rainbow passage (9). The interviewer posed a topic to the speaker, and the remaining time (generally twelve to thirteen minutes) was devoted to an extemporaneous monolog by the speaker. The interviewer responded briefly when appropriate, or when it was necessary to ask a new question for continuity.

A wide range of topics were covered, from describing a job to describing a frightening experience. Although one might argue that this approach in some sense constrained the data, casual listening of the recordings demonstrates that this is not the case. The topics generally provided a springboard for the speaker's thoughts, and the speech was usually conversatonal, fluent and quite varied. (With one subject, the suggested topic was consistently replaced by a wide variety of topics.)

Several observations should be noted which may be of considerable importance in practical situations. After the initial recording gain calibration for each session, no further gain adjustments were made. Subjects occasionally became bored or distracted, and either lowered their voice intensity or turned their heads away from the microphone. Conversely, subjects occasionally became intense on a topic and nearly "swallowed" the microphone, resulting in substantial low frequency waveform variablity due to breath bursts. Also, there was some stuttering, throat clearing, laughter, giggling and poor articulation.

In addition to these conditions, about half of the subjects acquired various degrees of colds during a two to three week period. All of these cases were recorded in the normal fashion, and no hand editing or deletion of any data was performed. The data used in this study consisted of only the extemporaneous speech material from the speakers, excluding the rainbow passage, word lists, etc. The total duration of the data base is 17 speakers x 10 sessions/speaker x approximately 13 minutes/session, or approximately 36.8 hours of data.

Several large population and long duration data bases have been reported in the literature (10,11). These were all text-dependent studies with short names or phrases. However, even the total duration of the large data base used by Das and Mohn is only one-tenth the total duration of the data base used in this study. The magnitude of this data base was extremely valuable for choosing feature subsets and defining reference sets which spanned varying periods of time.

Each audio tape was manually cued to the location where the extemporaneous portion of the interview began. Then real-time linear prediction analysis and disk storage of the analysis parameters was initiated. The data were low pass filtered at 3250 Hz and sampled at a 6500 Hz rate for compatibility with future applications to telephone systems and narrowband vocoder systems. The speech samples were preemphasized with a factor of 0.9, successive 128-point frames were multiplied by a Hamming window, and the autocorrelation method of linear prediction was used at a rate of fifty frames per second. The analysis was performed in real-time under Fortran control using a commercially available array processing system in conjunction with a PDP 11/45 computer (4,12). The analysis parameters for each speech frame were ten reflection coefficients, pitch period (obtained from a modified cepstral pitch tracker) and gain, and were stored in a quantized format of eight bits (one byte) per parameter. The process was terminated when the end of the tape was reached (defined as a thirty-second silence interval). The processing of each interview resulted in an analysis file of approximately 1000 disk blocks (512 bytes/block), and all interviews together required nearly half the total space of a 200-Mbyte disk (340,670 formatted disk blocks). In comparison, it would require ten 200-Mbyte disks to digitize all of the interviews with 12 bits/sample and to store directly without preprocessing.

Next, the analysis files were used to obtain long-term feature vectors, where each vector was the average of L_v successive voiced analysis frames. Unvoiced and silence frames were not included in this study, since it was felt that fundamental frequency was an essential speaker-dependent parameter. The vocoder analysis parameters consisted of fundamental frequency (the reciprocal of the pitch period), gain and ten reflection coefficients. For every interval L_v, long-term features based upon the mean, standard deviation and dispersion (standard deviation divided by mean) of the twelve parameters were computed, resulting in thirty-six-dimensional feature vectors. This feature set was defined in a reasonably general manner since analytic techniques for feature reduction may be used to find the most reasonable feature subsets for speaker recognition.

A summary of the number of feature vectors produced for all 170 interviews is given in Table 1. In this table, the data are partitioned into representative test and reference sets (13). Four choices of L_v were studied, namely $L_v = 30, 100, 300$ and 1000. The total number of feature vectors and the average real-time interval per feature vector as functions of L_v are also given.

TABLE 1 GOES HERE

It is important to consider the relationship between a particular value of L_v and the real-time interval of a long-term feature vector. Most significantly, a fixed number of voiced frames, rather than all of the voiced frames from a fixed elapsed-time interval, was chosen for analysis.

With extemporaneous speech, there may be intervals of ten to twenty seconds where very little or no voiced speech occurs (the speaker may pause, cough, laugh, etc.), leading to a variable voiced frame rate. If long-term features were a function of the voiced frame rate, then such features would not be reflective of only a speaker's speech sounds, but also his/her speech rate and style. While these additional characteristics might be a source of speaker-dependent information, they were not considered in this study, and consequently long-term features were made independent of the voiced frame rate.

The real-time interval for a long-term feature (seconds/feature) corresponds to a product of the following factors: 1) the number of voiced frames per feature vector (L_v), 2) the reciprocal of the voiced frame to total frame ratio (or the reciprocal of the voicing duty factor), and 3) the reciprocal of the number of analysis frames per second (or the reciprocal of the frame rate). In a previous study (5), the voicing threshold was set such that very smooth fundamental frequency (F_0) contours were observed on a real-time display system, and as a result, $L_v = 1000$ corresponded to approximately seventy seconds of real speech. For this study, the voicing threshold was determined by synthesizing the speech using the F_0 contour obtained, and then selecting the threshold that produced the subjectively best synthesis. The ear appears more sensitive to voiced speech segments which are synthesized as unvoiced, rather than the reverse, i.e. buzziness is typically preferred over whispery or hoarse speech. As a result, more voiced decisions were made, and $L_v = 1000$ in this study corresponded to approximately thirty-nine seconds of speech.

The feature vectors for each interview for each of the above values of L_v required approximately 301, 93, 33 and 13 disk blocks respectively, and a total of 74,800 disk blocks were required to store the feature vectors for the various L_v conditions for the 170 interviews. These data were then further processed as described in the next section.

III. Experiments in Parameter Variability

A. Intra-Speaker Variability

In a previous study (5), the within speaker (intra-speaker) variablity of the features for one male speaker was demonstrated to be a monotonically decreasing function of L_v from $L_v = 1$ to $L_v = 1000$ for a single fifteen minute session. Using the data base in this study, it was possible to study the intra-speaker variability for a larger number of male and female speakers, and in addition, it was possible to study the intra-speaker variability for cumulative sessions. If individual sessions are described by $S(i)$, $i=1,10$, then cumulative sessions may be described by $C(i)$, $i=1,10$, where $C(i) = S(1)+S(2)+...+S(i)$.

The standard deviations of the long-term averages of the fundamental frequency and the first reflection coefficient, denoted as $<F_0>$ and $<k_1>$ respectively, as measured over the cumulative sessions $C(i)$ for one male and one female speaker, are shown in Figure 1.

For both speakers and for each set of cumulative sessions, $\langle F_0 \rangle$ decreases as L_v increases. This behavior demonstrates that over long intervals, a speaker's average fundamental frequency is (probably) a good estimator of a characteristic or "habitual" value, and for successive long intervals, the deviation from the habitual value is small. For short intervals, influences such as speech prosody may mask the habitual value, and successive short intervals will deviate more widely from each other. This concept of habitual fundamental frequency is paralleled by the concept of habitual (perceived) pitch; the latter is used in speech therapy as a measure of acoustic improvement during treatment of a functional or organic voice disorder (14), and is an important factor in listener-based speaker recognition. For both speakers and for each value of L_v, there is a trend for $\langle F_0 \rangle$ to increase as more sessions are included (although there are exceptions, e.g. for the female speaker, $\langle F_0 \rangle$ for C(1) is greater than $\langle F_0 \rangle$ for C(2)). The dependence of $\langle F_0 \rangle$ on L_v can approximately be described as proportional to $L_v^{-1/2}$, which agrees with the theoretical relationship between the variance of a set of samples of a stationary random process, e.g., the L_v samples of F_0, and the variance of the process (5). In absolute terms, the standard deviation of the long-term fundamental frequency averages, over a time span of more than three months, varies from 17-23 Hz at $L_v = 30$ to 4-8 Hz at $L_v = 1000$ for the male speaker, and from 28-33 Hz at $L_v = 30$ to 8-11 Hz at $L_v = 1000$ for the female speaker.

The behavior of $\langle k_1 \rangle$ as L_v increases mirrors the behavior of $\langle F_0 \rangle$ as L_v increases. Since the value k_1 is a monotonic function of spectral slope of a first-order linear prediction inverse filter for speech (5,15), then a parallel explanation in terms of "habitual spectral slope" may be given, i.e., the longer the interval, the better the estimate of the habitual spectral slope. However, as more sessions are included, the behavior of $<k_1>$ differs from the behavior of $<F_0>$. For a given L_v, there is essentially no measurable increase in k_1 variability as the time period increases from one fifteen minute session to a period of nearly three months, with all ten sessions included. This trend is observed for the other speakers and the other long-term reflection coefficient averages, thus substantiating the presence of an "habitual spectral characteristic" for each speaker. Since the reflection coefficients are used to describe the vocal tract shape in an acoustic tube model (16), the result implies that the physical characteristics of a subject's vocal tract show no observable changes over at least several months.

Furui et al. (17-20) have examined speaker variability over intervals from a few weeks to several years. Their studies dealt with the variability of repeated word lists and short sentences. They found that for increasing time intervals from about three weeks to three months, spectral parameters such as reflection (PARCOR) or cepstral coefficients showed increasing variation. In contrast, the standard deviation of the reflection coefficients in this study show essentially no variation over time. Perhaps the data of Furui et al. were too linguistically constrained, and speakers never approached their habitual spectral characteristic.

In summary, inter-speaker variability based on averaged features decreases monotonically as the averaging interval increases. Furthermore, for a large averaging interval, inter-speaker feature variability is relatively consistent over a time period of three months. The next aspect of this study is a comparison which includes the intra-speaker information, e.g. a feature-by-feature analysis which uses the values of each feature from all subjects. If some features have small inter-speaker variance compared to the intra-speaker variance, then those features will not be useful for speaker recognition, and the performance of a classifier designed to recognize speakers from these features may be poor.

B. Variance Ratio Analysis

One method of measuring the usefulness of a feature for speaker recognition is the F-ratio or variance ratio (also referred to as the generalized Fisher discriminant) (7,10,19). The variance ratio of a feature is the quotient of the inter-speaker variance and the intra-speaker variance (11). In general, the larger the variance ratio for a particular feature, the greater the probable contribution of the feature in distinguishing the speakers (13), but this property is strongly dependent on the data and the experimental procedure. However, the variance ratio does not account for inter-feature correlations, and if two features with high variance ratios are highly correlated, then the inclusion of both parameters might be somewhat redundant (7).

1. Trends as a function of population

The variance ratios for the case $L_v = 1000$ and cumulative sessions 1-10 are shown in Figure 2 for the male and female speakers separately, and in Figure 3 for two subsets of the male speakers. Only the variance ratios of the mean and standard deviation features are shown. The variance ratios of the dispersion features were consistently low, and therefore believed to contribute very little toward speaker recognition in this study.

There are noticable differences in the variance ratios between the male and female populations. Based on relative magnitudes, the features $< (k_9)>$, $< (k_8)>$ and $<k_1>$ would be the most significant for identifying the male population, while $< (k_7)>$, $< (k_8>$ and $<k_8>$ would be the most significant for identifying the female population. If the male population is arbitrarily divided into two equal-sized subsets, there are pronounced changes in the variance ratios. For the first set of male speakers, $<k_1>$, $<F_0>$ and $<k_2>$ have the largest variance ratios, and for the second set of male speakers, $<k_4>$, $<k_8>$ and $<k_3>$ have the largest variance ratios. These results show the need to have a substantially larger speaker population in order to characterize the parameters of major importance. However, it is estimated that to obtain variance ratios which would exhibit consistent trends for a set of speakers and for subsets of the speakers, a much larger data base, possibly more than 100 speakers, would be required.

In the previous paper (5), for a smaller and more homogeneous data base, $<k_2>$ and $<k_6>$ were found to be the most significant parameters. These large variance ratios would be physical evidence for the importance of the first and third formants in voiced speech (5). This larger population base, however, shows no such relationships. The conclusion is that for studies with linguistically unconstrained speech, parameter ranking using variance ratios should be used cautiously. The parameters with large variance ratios may change depending on how the data are partitioned, and the features with small variance ratios may be important for achieving good speaker recognition if the data partitioning is changed. (Conversely, it will be shown that some parameters with small variance ratios may actually degrade speaker recognition.)

2. Trends as a function of L_v and time-spacing

The variance ratios were determined for the case $L_v = 100$ and cumulative sessions 1-10 (Figure 4), and for the case $L_v = 1000$ and cumulative sessions 1-2 (Figure 5). Comparing Figures 2 and 4, which only differ by the averaging interval L_v, the variance ratios generally maintain the same relative relationships, i.e. the features which have the relatively larger variance ratios for $L_v = 1000$ also have the relatively larger variance ratios for $L_v = 100$. However, the absolute values of the variance ratios are smaller for $L_v = 100$ than for $L_v = 1000$. Comparing Figures 2 and 5, which only differ by the number of sessions, the relative relationships and the absolute values of the variance ratios are similar for two cumulative sessions and for ten cumulative sessions. However, a slight decrease in the absolute values of the variance ratios for ten cumulative sessions is observed. If the inter-speaker variance is assumed relatively constant for two or ten cumulative sessions, then the slight decrease in variance ratios for ten sessions over two sessions correlates with the slight increase in standard deviations observed in Figure 1. This result further establishes that a speaker's habitual features, when measured over a relatively long interval (greater than thirty seconds), do not show appreciable changes over time periods up to three months.

3. Further observations

It is also evident that the variance ratios for the mean features generally have larger values than the corresponding variance ratios for the standard deviation features. The variance ratios for the dispersion features are in turn substantially lower in value than the corresponding variance ratios for the standard deviation features. Features based upon gain have consistently small variance ratios.

IV. Speaker Recognition

Speaker recognition was based on a weighted Euclidean distance metric (5,7,11), where the mean vector and inverse covariance matrix for each of the seventeen speakers were estimated from feature vectors in the specified reference set. All thirty-six dimensions were used initially. The distances between each reference class and each test vector were computed, and the test vector was assigned to the reference class which yielded the smallest distance. For speaker identification, a tally was taken of the number of correct choices. For speaker verification, the distances were stored for further analysis with a variable distance threshold. The method of cross-validation in both directions was used (11), where independent subsets of the data were cyclically treated as test and reference groups, and the speaker recognition scores for each cycle were averaged for the final scores.

Atal (7) and Bricker *et al.* (10) discussed three possible choices for a distance metric. Each metric was a positive semidefinite form which could be described by $d = (X-Y_i)\ \underline{M}\ (X-Y_i)^T$, where X was a vector to be classified, Y_i was the mean vector for class i, and $\underline{M}$ was a weighting matrix. The choices for $\underline{M}$ were a pooled covariance matrix $\underline{W}^{-1}$ from all speakers, an individual covariance matrix $\underline{W}_i^{-1}$ from each speaker, or a discriminant matrix $\underline{D}$ composed of the eigenvectors of $\underline{W}^{-1}\ \underline{B}$, where $\underline{B}$ was the between-class covariance matrix.

The use of the discriminant matrix $\underline{D}$ requires sufficient knowledge of the inter-speaker variability, which may be difficult to attain unless an extremely large number of speakers is used. Atal (7) and Bricker *et al.* (10) preferred the pooled covariance matrix $\underline{W}$ over the individual covariance matrix $\underline{W}^{-1}$. Their rationale was that data limitations (less samples than dimensions) frequently result in a singular (noninvertible) covariance matrix, and that one pooled covariance matrix would adequately represent all speakers, even though speaker dependent data is contained in individual covariance matrices and subsequently is not used.

From Table 1, the average number of feature vectors per speaker per session for L_v = 30,100,300,1000 is 685 (116411/170), 205 (34862/170), 68 (11563/170) and 20 (3413/170) respectively. For L_v = 30,100 or 300, with thirty-six dimensions, the individual covariance matrices were never singular for any number of pooled sessions. For L_v = 1000, with thirty-six dimensions, the individual covariance matrices were singular if less than three sessions are pooled. Furthermore, Kanal (14) has suggested that ten times the number of dimensions is an adequate sample size for good covariance matrix estimates with normal probability distribution assumptions. For five sessions and thirty-six dimensions in a reference class, the factors which relate sample size to dimensionality for L_v = 30,100,300,1000 are 95 (685*5/36), 28 (205*5/36), 9 (68*5/36) and 3 (20*5/36) respectively. For L_v = 1000, sessions as long as forty-five minutes would have been necessary to produce a factor near ten, but a factor as large as ten is probably not needed for features which are themselves the average of 1000 frames of data. However, fifteen minutes was a sufficient duration for the other values of L_v, as well as an upper limit of endurance for the subject and interviewer. It was felt that the advantages gained through the use of individual covariance matrices outweighed potential problems of undersampling the speaker's statistics. In a practical situation, relatively long sessions would be necessary for sufficient accumulation of speaker's reference data, but thereafter the speaker could be verified approximately every thirty-nine seconds.

A. Trends as a function of L_v

For the first series of tests, the first five sessions were treated as the reference data, the second five sessions were treated as the test data, and then vice-versa. Results are shown in Table 2. In Table 2a, it is seen that the average scores for the probability of correct identification P(CI) monotonically increase from 60% to nearly 92% as L_v increases from 30 to 1000 respectively. A confusion matrix of identification errors shows that no one speaker is more difficult to identify than any other speaker. In Table 2b, as L_v increases, the speaker verification equal error probability (probability of false acceptance P(FA) equals the probability of ıalse rejection P(FR)) monotonically decreases from 43.1% to 8.8%. This trend is principally due to the P(FA) behavior, since the P(FR) behavior does not change appreciably with L_v (5). Although the distance threshold for a given probability of correct acceptance and fixed dimensionality (under multivariate normal assumptions) may be analytically obtained, the distance threshold for the equal error probability can only be determined experimentally. In Table 2c, the equal error probability distance threshold is seen to monotonically increase as L_v increases.

TABLE 2 GOES HERE

It is interesting to illustrate the difference between text-independent speaker recognition with and without linguistic constraints. Sambur has proposed an orthogonal linear prediction set of parameters for text-independent speaker recognition (4). Within the context of a linguistically constrained experiment where all speakers spoke the same set of sentences, Sambur's text-independent results (in the sense that the reference sentences were different from the test sentences) were near 94%. The orthogonal linear prediction parameters are essentially equivalent to a linear transformation of the long-term reflection coefficients averages used in this study if $L_v = 1$ (equivalent to no averaging). If all linguistic constraints are removed, and if little or no averaging is used, the results of Table 2 indicate that the speaker identification scores for a true text-independent situation with a reasonable number of speakers will be quite poor (even for $L_v = 30$, P(CI) is bounded from above at 62%). A similar statement follows for the case of speaker verification.

B. Trends as a function of time spacing

Rosenberg (21) has noted that one of the most important considerations in designing a data base is the time period over which utterances are collected and the methods for establishing reference patterns over time. Following the pictorial scheme of Furui et al. (14-17) for illustrating reference and test sets over time, speaker recognition for four cases shown in Figure 6 were investigated. Reference sets were composed of from two to five successive sessions (with a time interval of at least one week between sessions). No comingling such as odd-numbered reference sessions and even-numbered test sessions was allowed. For each case, the reference and test sets were composed of equal numbers of successive independent sessions, and two-direction recognition tests (as described above) were made for the four L_v cases.

The results are presented in Table 3. It is seen that for all L_v conditions, higher scores were obtained as the number of cumulative sessions increased.

TABLE 3 GOES HERE

The differences in the speaker identification score between the first two sessions and the first five sessions is around 15% for all L_v cases shown (L_v = 1000 was not used for two sessions since the covariance matrices were singular). It is interesting to note that in a text-dependent speaker verification experiment with different parameters and approaches, Luck (22) found that speech samples collected over a five week period gave the best results.

C. Trends as a function of feature subsets

In a previous section, it was noted that the dispersion features had very small variance ratios, whereas the mean features as a group consistently had the largest variance ratios. How would recognition scores compare if the dispersion features were omitted, or if only the mean features were included? The recognition test for L_V = 1000 and five sessions per reference and test set was repeated using several different feature subsets, based on an analysis of the magnitudes of the variance ratios. In one case, only the twelve mean features were used, and in a second case, only the twenty-four mean and standard deviation features were used. The average scores for the two cases were P(CI) = 93.6% with P(FA) = P(FR) = 14.5%, and P(CI) = 96.8% with P(FA) = P(FR) = 7.2% respectively. For comparison, the comparable average scores for all thirty-six features (Table 2) were P(CI) = 91.6% with P(FA) = P(FR) = 8.8%.

Not only did both of these new cases based on feature subsets yield better scores than the original thirty-six dimension feature set, but in the second case, the identification score was markedly increased by more than 5%. This result is a significant practical illustration that the inclusion of some parameters which would hopefully improve performance (or at worst case would have no effect on performance), can sometimes actually degrade the system performance in an open test. In a closed test with the distance metric used in this study, where a reference set also is used as a test set, this theoretically cannot happen. Closed tests on this data base verified that monotonic increases in the number of features produced monotonic increases in the P(CI) and monotonic decreases in equal error probability P(FA) = P(FR).

This improved performance by eliminating features with relatively small variance ratios was the basis for one additional test with a feature subset. In considering the remaining twenty-four features, the gain-related features had very small variance ratios, and furthermore, the inclusion of gain-related features was difficult to physically justify. In fact, it could be argued that even if these features helped, they should not be included because they may simply reflect a speaker's position, interest, etc. during the recording session. Therefore, the recognition test with only twenty-four features was repeated with the gain-related features removed, and the performance of this last test with only twenty-two parameters was better than any previous test. The final results of this study using only the twenty-two fundamental frequency and reflection coefficients long-term averages are shown in Table 4.

TABLE 4 GOES HERE

These results are extremely promising for future studies in many areas of speaker recognition. This substantially large testing effort (over eighty million distance measurements) has shown that realistic and acceptable speaker identification and speaker verification can be achieved with text-independent linguistically unconstrained speech.

The cumulative probability functions (Figure 7A) and the probability density functions (Figure 7B) for false rejection and false acceptance may be used to compare the inter- and intra-speaker distances in the verification task. These curves are derived from the first half of the final speaker verification test with 22 features, $L_v = 1000$, reference sessions 1-5 and test sessions 6-10. The equal error point is graphically depicted as the crossover point of the two cumulative probability curves in Figure 7A. This equal error point is found at a distance threshold where the probability of false acceptance (i.e. acceptance of an imposter) is equal to the probability of false rejection (i.e. rejection of a correct speaker).

The probability density functions (pdfs) in Figure 7B show the distribution of the intra- and inter-speaker distances. The crossover point in Figure 7A divides each of the pdfs into two sections, with the area under the intra-speaker pdf to the right of the dividing line equal to the area under the inter-speaker pdf to the left of the dividing line. For this data, the equal error crossover point is close to the intersection of the two pdfs, but only identical and symmetric pdfs will always have identical crossover and intersection points.

For test sessions 6-10 with $L_v = 1000$, there were a total of 1708 test vectors from the 17 speakers. The distances between each of these test vectors and the correct reference speaker comprise the intra-speaker distance space. A histogram of these intra-speaker distances is shown in Figure 8A. The mean and standard deviation of the histogram distances were used to approximate normal and log-normal distributions. For the open test, there is no underlying theoretical distribution, and a chi-square test was used to measure the goodness of fit of the normal and log-normal distributions. The log-normal distribution had the smallest chi-square measure. Analogously, the distances between each of the 1708 test vectors and each of the 16 incorrect speakers (i.e. eliminating the reference speaker who is a correct match to the test vector) comprise the inter-speaker distance space. A histogram of the 27,318 inter-speaker distances is shown in Figure 8B. A log-normal distribution is a better fit to the inter-speaker histogram than a normal distribution, but not as good a fit as with the intra-speaker histogram.

V. Summary

The significance and value of long-term feature averaging for text-independent speaker recognition with linguistically unconstrained speech has been demonstrated. This study used practical analysis conditions of telephone-range spectral width (0-3250 Hz) and parameters obtained from a linear prediction vocoder. All parameter-related computations were performed in real time using 16-bit integer arithmetic, and all parameters were further quantized into an 8-bit format for efficient disk storage.

The recording environment was controlled by recording the speakers with a condenser microphone in an IAC sound room. An important extension of this work would be to reprocess the "clean-text" audio tapes through various channel disturbances such as the telephone system to determine the robustness of the approach in less ideal environmental conditions (20). Also, in some situations, reference data may be obtained in a clean environment and subsequent speaker recognition attempted in a noisy environment. This area also requires investigation.

Although seventeen speakers is not a trivial population size, it appears that for determining the importance of individual features for speaker recognition using linguistically unconstrained text, a substantially larger population base is required. It was found that features obtained from only one or two sessions of a given population are relatively unchanged over a much larger number of time-spaced sessions, where there was at least one week between sessions. Other features should also be investigated. It has been suggested that mean deviations (21) may prove more useful than the standard deviations used in this study. Further research is also required to assess the conditions, e.g., the number of long-term samples from a speaker, for obtaining a good estimate of the mean and variance of a speaker's characteristics.

An assumption throughout has been that only voiced speech frames are to be used in the analysis. If this assumption was not necessary, or if only slight degradation occurred if both voiced and unvoiced speech frames were included, the process would be simplified computationally, and in addition, 1000 frames per average would correspond to a real time interval only about half as long as required here.

The best speaker recognition was obtained when 1) five sessions successively separated by at least one week were used to define the reference set, 2) the mean and standard deviation of the long-term averages of the fundamental frequency and reflection coefficients were used, and 3) each feature was obtained by averaging 1000 voiced analysis frames (corresponding to average real-time intervals of about thirty-nine seconds). With approximately eighteen hours of reference data and eighteen hours of independent test data from seventeen speakers, spaced over nearly three months in time, an average speaker identification score of 98.05% and an average equal error speaker verification rate of 4.25% were measured.

Acknowledgements

The authors would like to acknowledge the substantional contributions of the following people on this project: Dr. Beatrice Oshika for organization of the data base, and Rob Arnott and Ted Applebaum for programming assistance with the real-time system and the non-real-time batch processing programs. Also, the authors appreciate the helpful comments from Dr. Jared Wolf and Dr. Melvyn Hunt on an early draft of this paper.

References

1) B.S. Atal, Effectiveness of linear prediction characteristics of the speech waves for automatic speaker identification and verification, *The Journal of the Acoustical Society of America*, vol. 55, pp. 1304-1312, 1974.

2) K.P. Li and G.W. Walker, Talker differences as they appear in correlation matrices of continuous speech spectra, *The Journal of the Acoustical Society of America*, vol. 55, pp. 833-837, 1974.

3) K.O. Mead, Identification of speakers from fundamental frequency contours in conversational speech, *Joint Speech Research Unit*, Report 1002, 1974.

4) M.R. Sambur, Speaker recognition using orthogonal linear prediction, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, pp. 283-289, 1976.

5) J.D. Markel, B.T. Oshika and A.H. Gray, Jr., Long-term feature averaging for speaker recognition, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-25, pp. 330-337, 1977.

6) M.J. Hunt, J.N. Yates and J.S. Bridle, Automatic speaker recognition for use of communication channels, *IEEE ICASSP Conference Record*, 764, 1977.

7) B.S. Atal, Automatic recognition of speakers from their voices, *Proceedings of the IEEE*, vol. 64, pp. 460-475, 1976.

8) B. Beek, E.P. Newberg, and D.C. Hodge, An assessment of the technology of automatic speech recognition for military applications, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-25, pp. 310-322, 1977.

9) G. Fairbanks, *Voice and Articulation Handbook*, Harper, New York, 1960.

10) P.D. Bricker, R. Gnanadesikan, M.V. Mathews, S. Pruzansky, P.A. Tukey, K.W. Wachter and J.L. Warner, Statistical Techniques for talker identification, *Bell Systems Technical Journal*, vol. 50, pp. 1427-1454, 1971.

11) S.K. Das and W.S. Mohn, A scheme for speech processing in automatic speaker verification, *IEEE Transactions on Audio and Electroacoustics*, vol. AU-19, pp. 32-43, 1971.

12) R.D. Arnott and J.D. Markel, Fortran control of real-time signal processing, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, to be published, August, 1978.

13) L. Kanal, Patterns in Pattern Recognition: 1968-1974, *IEEE Transactions on Information Theory*, vol. IT-20, pp. 697-722, 1974.

14) L. Travis, Handbook of Speech Pathology and Audiology, Appleton Century Croft, New York, 1971.

15) A.H. Gray, Jr. and J.D. Markel, A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-22, pp. 207-217, 1974.

16) H. Wakita, Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms, IEEE Transactions on Audio and Electroacoustics, vol. AU-21, pp. 417-427, 1973.

17) S. Furui and F. Itakura, Talker recognition by statistical features of speech, Electronic Communications of Japan, vol. 56-A, pp. 62-71, 1973.

18) S. Furui, F. Itakura, and S. Saito, Talker recognition by longtime averaged speech spectrum, Electronic Communications of Japan, vol. 55-A, pp. 54-61, 1972.

19) S. Furui, An analysis of long-term variations of feature parameters of speech and its application to talker recognition, Electronic Communications of Japan, vol. 57-A, 1974.

20) S. Furui, F. Itakura and S. Saito, Personal information in the long-time averaged speech spectrum, Review of the Electrical Communication

Laboratories, vol. 23, pp. 1133-1141, 1975.

21) A.E. Rosenberg, Automatic speaker verification: a review, Proceedings of the IEEE, vol. 64, pp. 475-487, 1976.

22) J.E. Luck, Automatic speaker verification using cepstral measurements, The Journal of the Acoustical Society of America, vol. 46, pp. 1026-1031, 1969.

FIGURES

TABLES

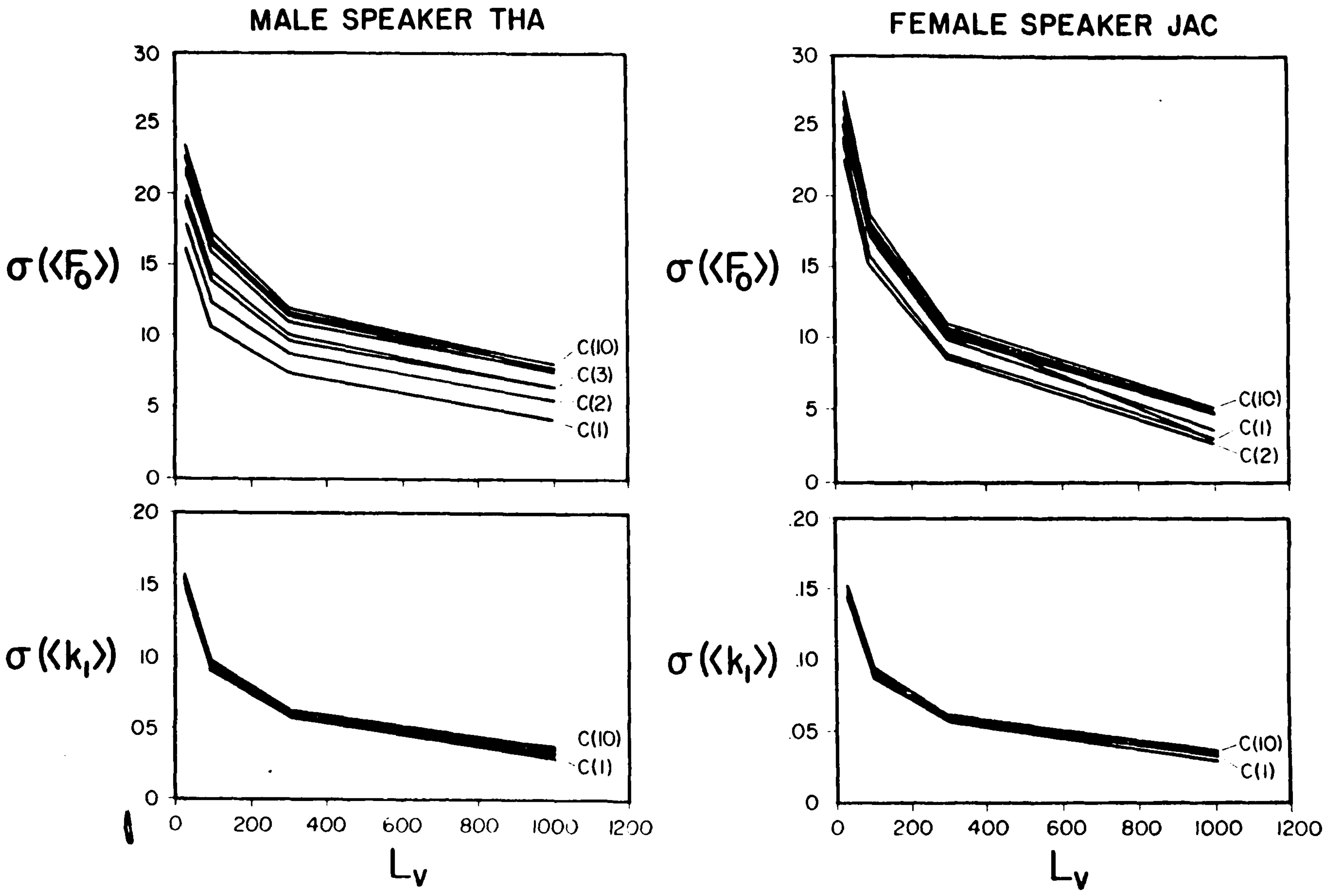


Fig. 1 Standard deviation of long-term features as a function of L_v, the number of voiced frames per feature vector.

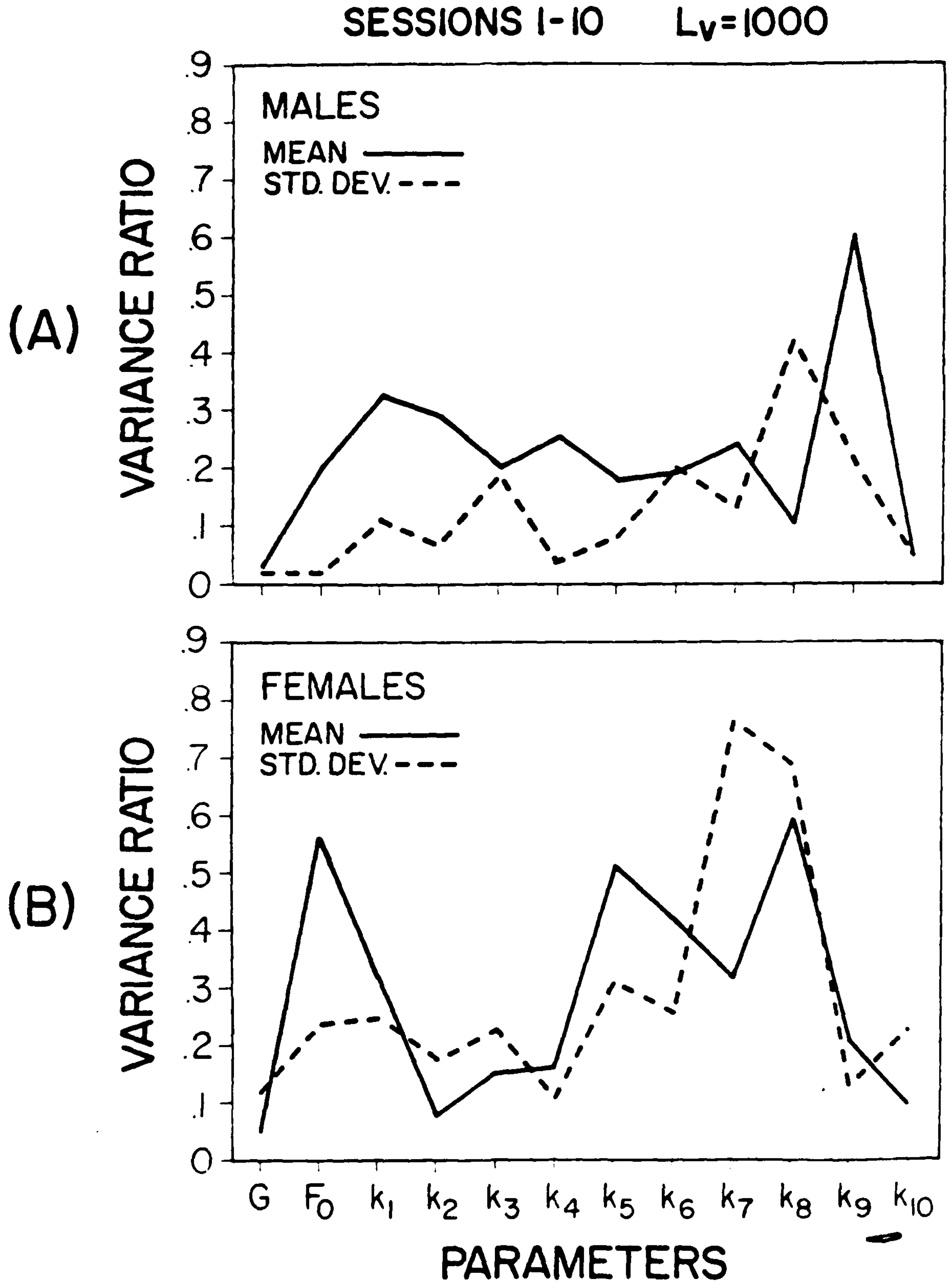


Fig. 2 Variance ratios from all 10 sessions as a function of long-term mean and standard deviations of parameters.
A) all male speakers
B) all female speakers. L_v=1000

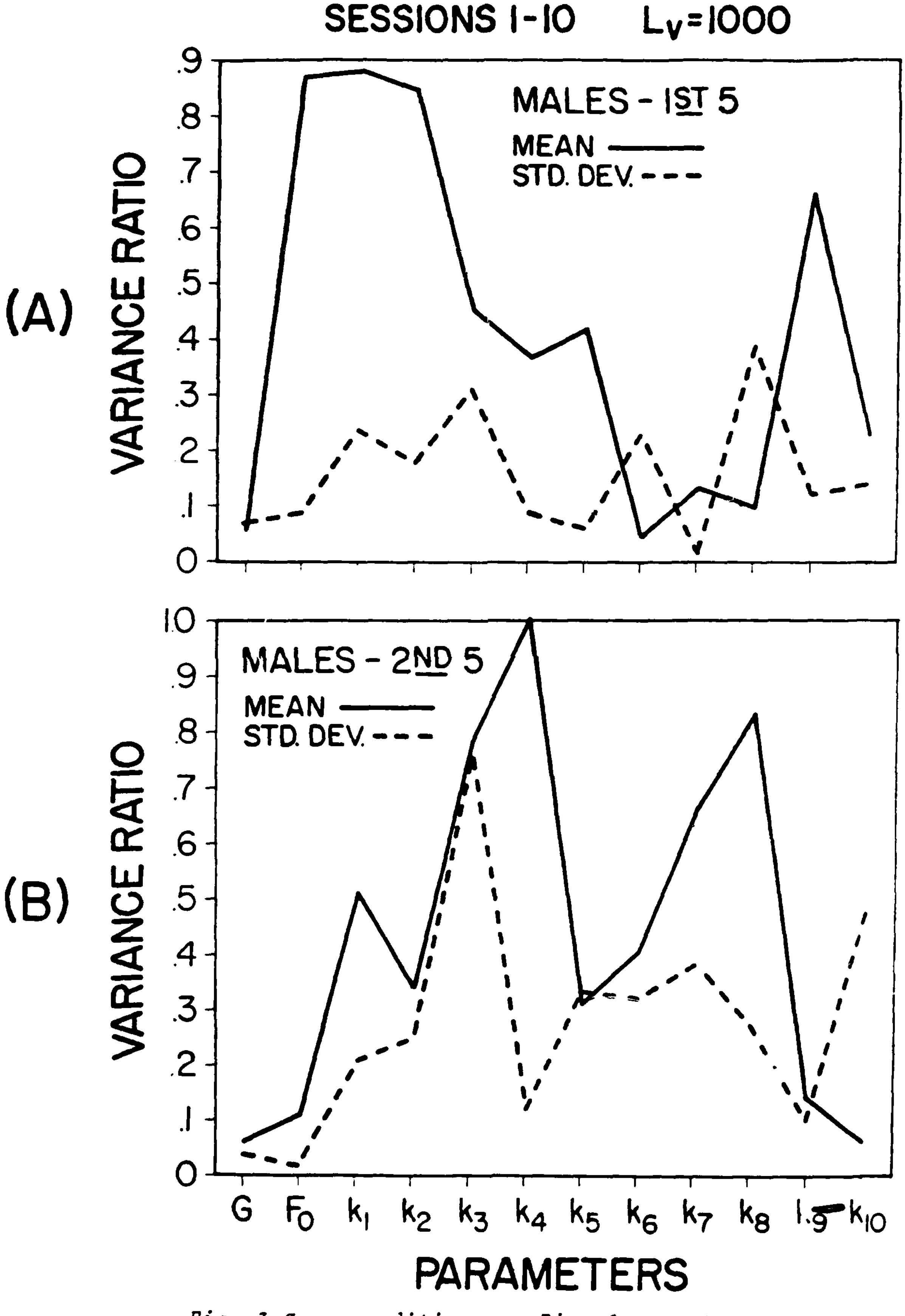


Fig. 3 Same conditions as Fig. 2 except
A)male speakers:first five
B)male speakers:second five

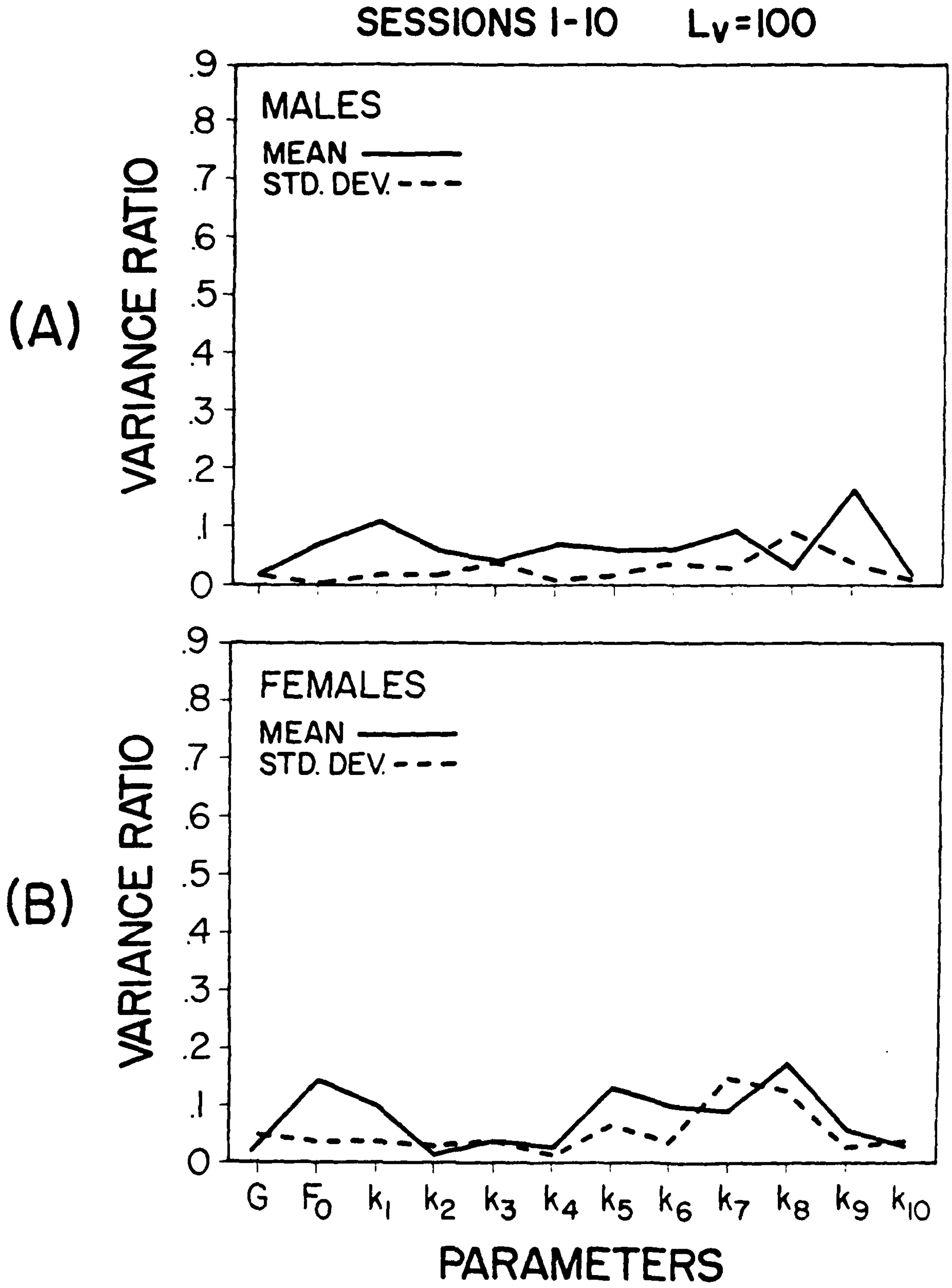


Fig. 4 Same conditions as Fig. 2 except that L_v=100:
A) all male speakers
B) all female speakers

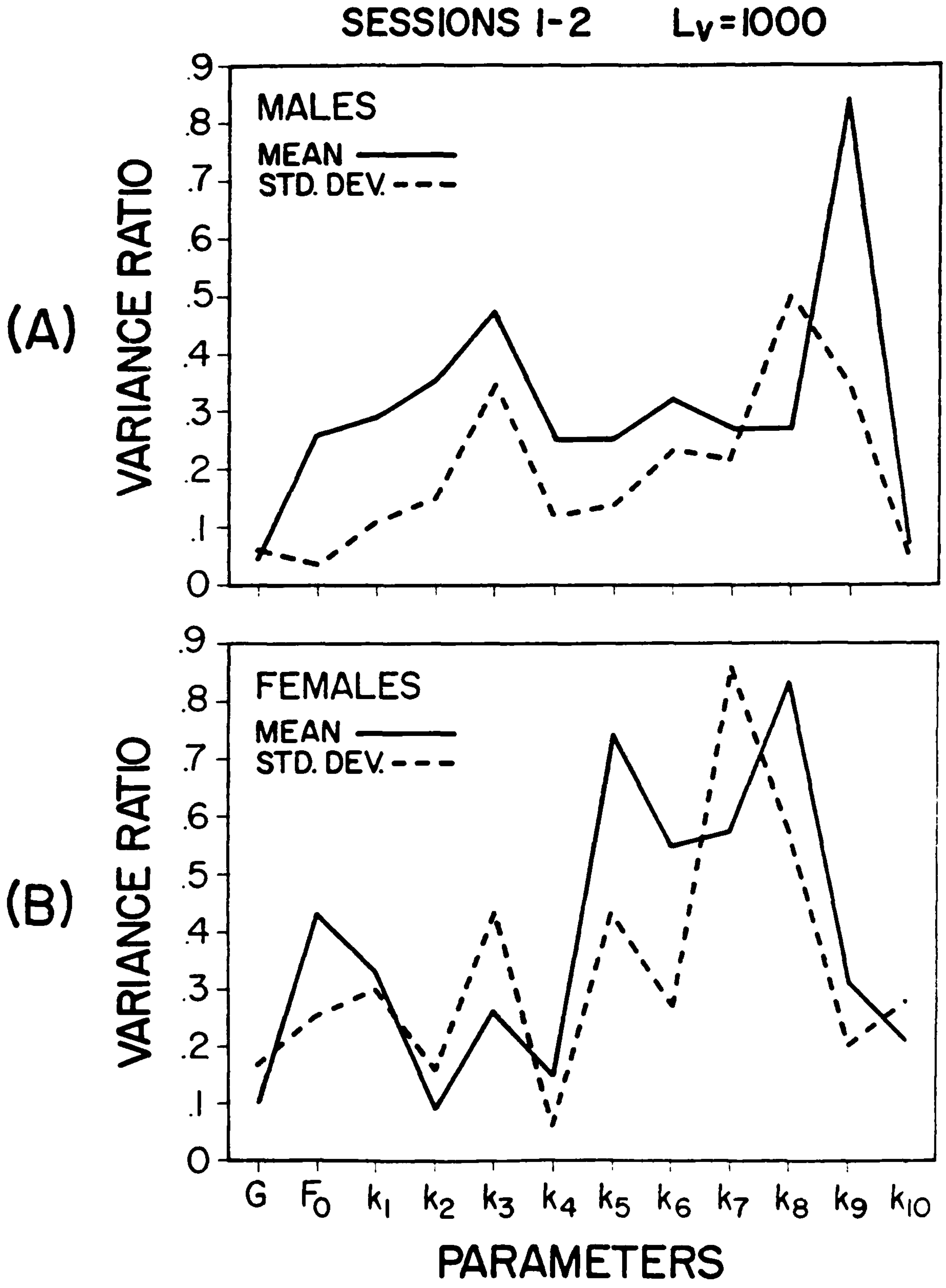


Fig. 5 Same conditions as Fig. 2 except only sessions 1-2 shown:
A) all male speakers
B) all female speakers

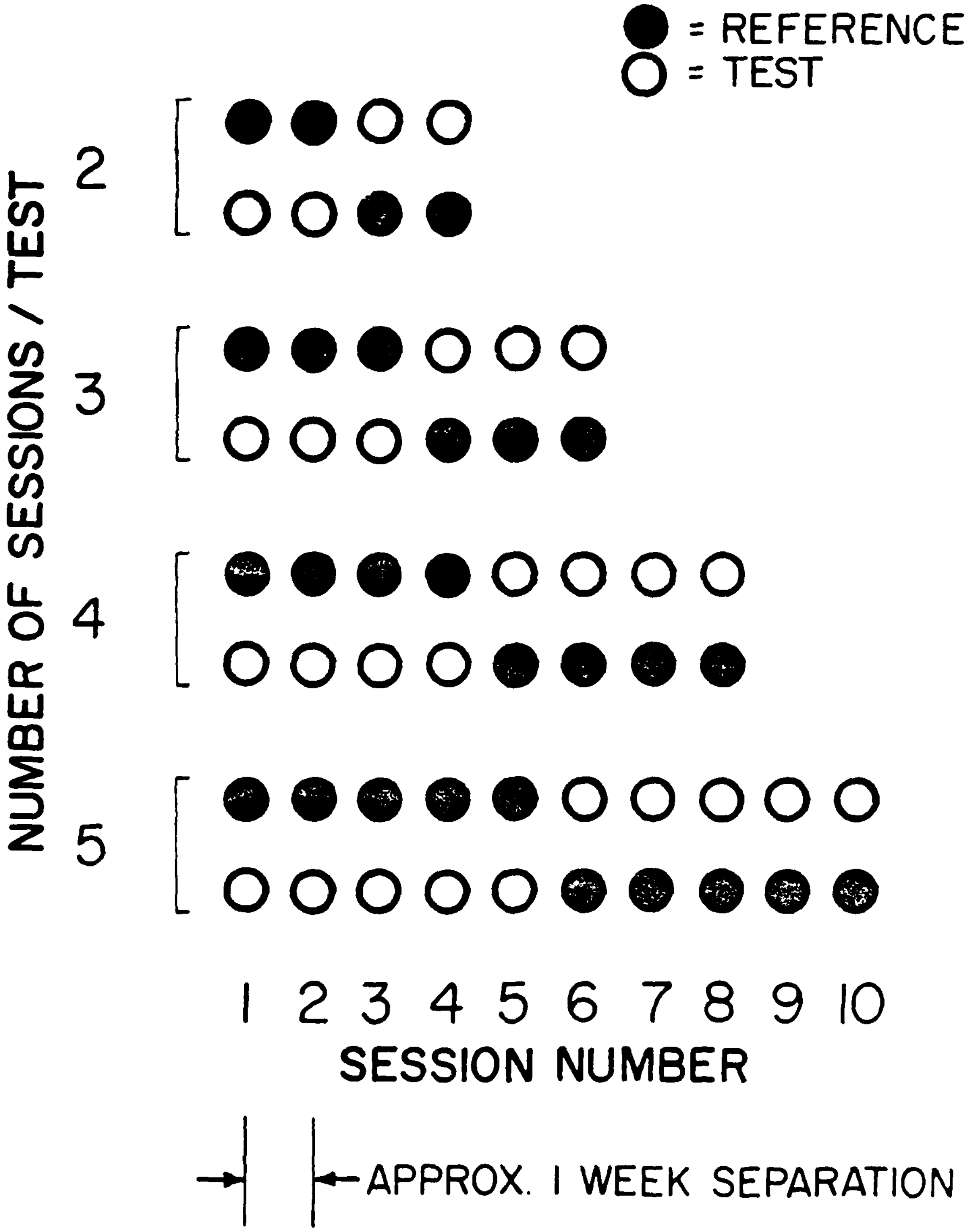


Fig. 6 Relations between reference samples and test samples for experimental results of Table 3.

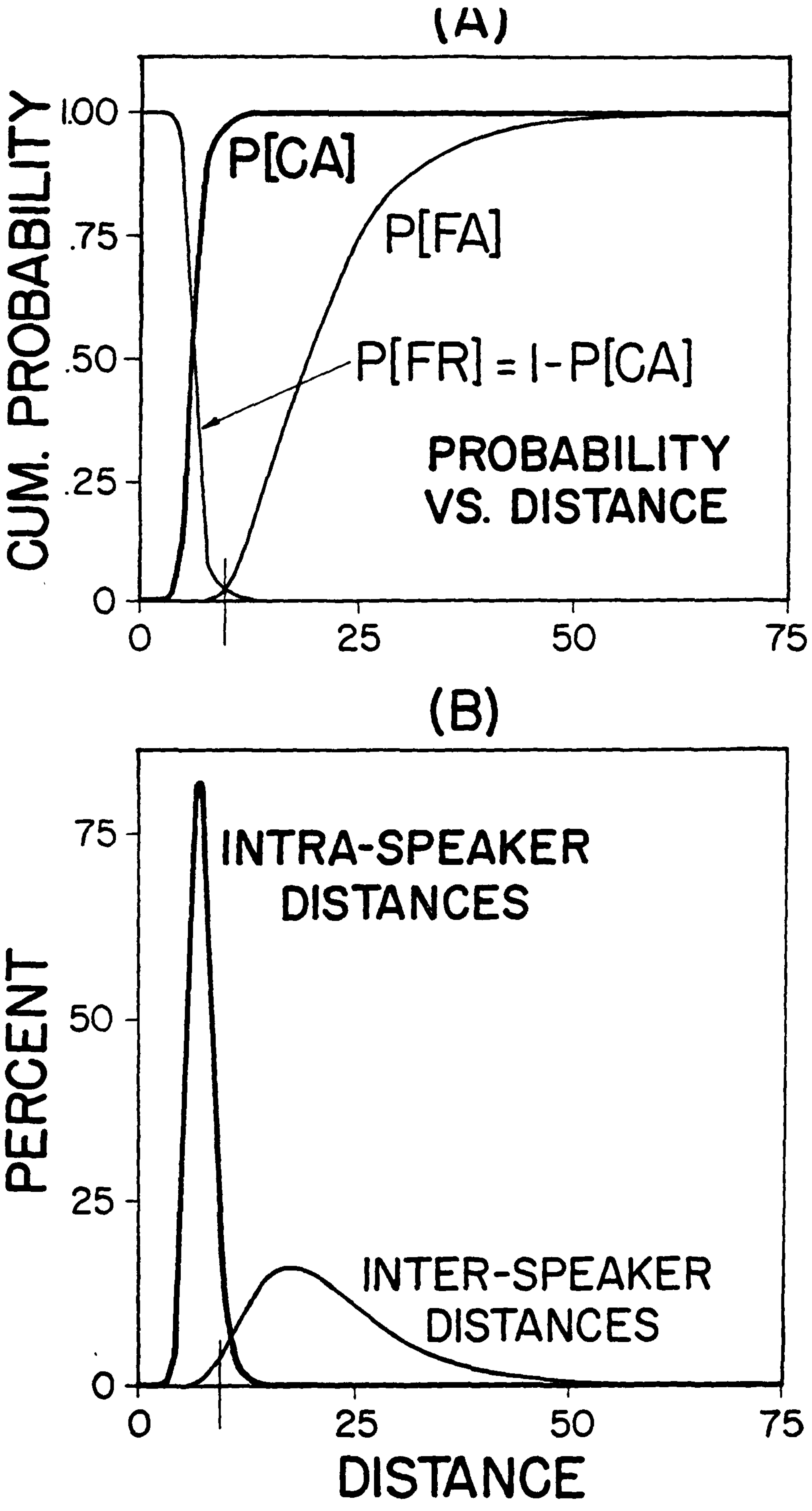


Fig. 7 Intra and inter- Speaker Comparisions
A) Cumulative Probability
B) Probability Density Estimates

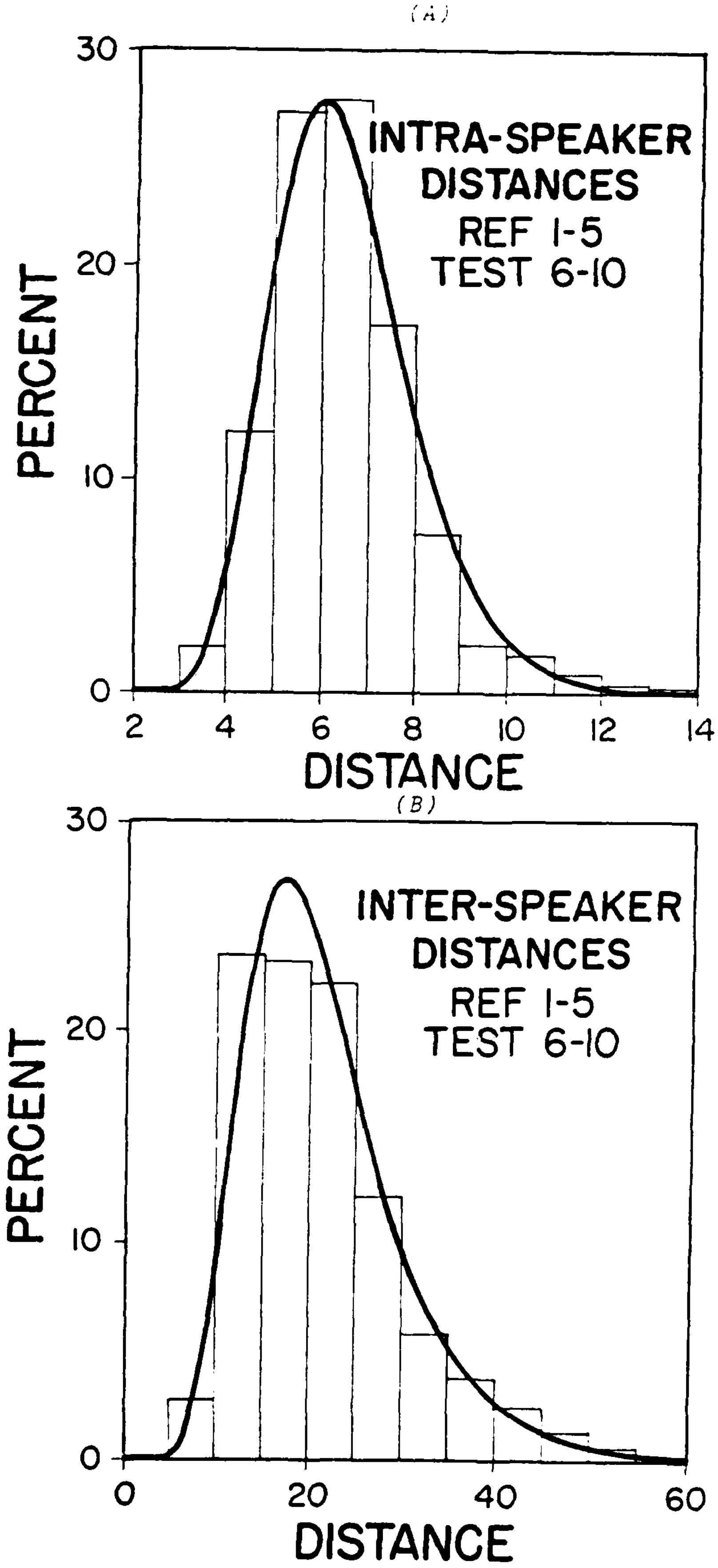


Fig. 8 Distance Histograms and Models
A)Intra-speaker Distances
B)Inter-speaker Distances

		L_V			
		30	100	300	1000
REFERENCE SESSION	1-5	58,379	17,486	5,799	1,712
	6-10	58,032	17,376	5,764	1,701
TOTAL NUMBER OF TOKENS		116,411	34,862	11,563	3,413
AVERAGE REAL TIME INTERVAL (SEC)		1.14	3.80	11.47	38.85

Table 1. Number of feature vectors and average real-time interval (RTI) for each L_v condition.

(A)

SPEAKER IDENTIFICATION

Percent of correct choices based on minimum distance

SESSION		L_V			
REF.	TEST	30	100	300	1000
1-5	6-10	61.20	78.65	88.20	93.34
6-10	1-5	59.87	75.48	85.27	89.77
AVERAGE		60.54	77.06	86.74	91.56

(B)

SPEAKER VERIFICATION

Percent of false acceptances and false rejections based
on equal error criterion

SESSION		L_V			
REF.	TEST	30	100	300	1000
1-5	6-10	43.4	27.8	10.7	9.4
6-10	1-5	42.8	26.9	10.5	8.2
AVERAGE		43.1	27.4	10.6	8.8

(C)

SPEAKER VERIFICATION

Threshold distance based on equal error criterion

SESSION		L_V			
REF.	TEST	30	100	300	1000
1-5	6-10	5.79	7.52	9.78	18.84
6-10	1-5	5.85	7.58	10.85	21.10
AVERAGE		5.82	7.55	10.32	19.97

Table 2. Speaker recognition based on partitioning data in half and with 36 long-term features

SESSIONS			L_V			
NO.	REF.	TEST	30	100	300	1000
2	1-2	3-4	50.36	64.34	71.18	—
2	3-4	1-2	53.45	67.95	75.31	—
3	1-3	4-6	54.29	70.03	79.12	80.58
3	4-6	1-3	57.04	72.69	82.14	89.30
4	1-4	5-8	59.91	76.41	86.73	92.85
4	5-8	1-4	59.26	74.62	83.45	86.34
5	1-5	6-10	61.20	78.65	88.20	93.34
5	6-10	1-5	59.87	75.48	85.27	89.77

Table 3. Percent of speakers correctly identified as a function of the number of reference sessions

FINAL RESULTS OF 2-WAY TESTING ON 38 HOURS OF EXTEMPORANEOUS SPEECH

SESSION NUMBERS		SPEAKER IDENTIFICATION	SPEAKER VERIFICATION
REF.	TEST	P(CI) (%)	P(FA) = P(FR) (%)
1-5	6-10	98.65	3.3
6-10	1-5	97.45	5.2
AVERAGE SCORE		98.05	4.25

Table 4. Performance with fundamental frequency and reflection coefficient mean and standard deviation long-term features, L_v = 1000 (average real-time interval = 39 seconds).